

# **Sentiment analysis on students' real-time feedback**



**Nabeela Altrabsheh**

The thesis is submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy of the University of Portsmouth

February 2016



## **Abstract**

Previous literature identifies that students' real-time feedback is important in the learning process. There are numerous studies that have collected students' feedback in real time. However, they include several limitations of which the most important is analysing the feedback. In this thesis, we address these limitations by proposing a system that will automatically analyse students' feedback in real time and present the analysis results to the lecturer. To create such a system, we propose the use of sentiment analysis.

The extensive literature highlights the importance of this research in the sentiment analysis field, as there is no research exploring sentiment analysis for students' real-time feedback and research in the educational domain has been focused mainly on e-learning. The literature shows that emotions are also important in learning and could be detected using sentiment analysis. Consequently, this research is also important as there is little research in detecting emotion from students' feedback.

Polarity detection is explored and two experiments were done to identify the best combination of preprocessing, features and machine learning techniques to create an optimal model for polarity detection of students' feedback. As a result, we found that the optimal model was to use a lower level of preprocessing, unigrams for features and Complement Naive Bayes for the classifier.

Emotion detection in students' feedback is also explored. Two experiments were done to find the optimal model to detect emotions related to learning from students' feedback. The results showed that models which detect a single emotion have a better performance than multiple emotion models. Three emotions (i.e. Amused, Bored, and Excited) were more easily detected than others. The optimal model to detect emotion included a low level preprocessing, unigrams as features and Complement Naive Bayes as the classifier.

Sarcasm detection in students' feedback is explored. Our results showed that the lower level of preprocessing led to the best performance. Moreover, by adding other features, such as polarity and emotions, to the unigrams, sarcasm detection increased. Lastly, the best classifier to detect sarcasm was Complement Naive Bayes.

Visualisation of the sentiment analysis models results is also investigated. Six visualisations were created and evaluated by lecturers. The findings indicated that there lecturers had no strong preference for a specific visualisation.

The system evaluation was performed in real-life settings. As a result, we found that the lecturers were highly satisfied with the system, while the students had a neutral towards negative view of it.

## **Acknowledgements**

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to do without the support and guidance that I received from many people.

Firstly, I would like to express my sincere gratitude and thanks to my advisor Dr. Mihaela Cocea for the continuous support of my PhD study, for her patience, motivation and immense knowledge. Her guidance, insightful comments and encouragement helped me in all the time of research and writing of this thesis. I have been extremely lucky to have a supervisor who cared so much about my work and who responded to my questions and queries very promptly. Without her support, it would not be possible to conduct this PhD. I could not have imagined having a better supervisor for my PhD.

I also would like to thank Dr. Sanaz Fallahkhair for her generous time and valuable help in this research. Her valuable insights and feedback greatly shaped my thesis. She has been actively interested in my work and has always been available to advise me.

Furthermore, I would like to thank the School of Computing, not only for providing me with funding which allowed me to attend conferences, but also for giving me the opportunity to meet many interesting people.

I want to use this opportunity to express my deep and sincere gratitude to Dr. Khaldoon Dhou for his aspiring guidance, invaluable constructive criticism and friendly advice during this PhD. I am sincerely grateful to him for sharing his truthful and illuminating views on a number of issues related to the PhD. I would also like to thank him for his endless support and encouragement.

Finally, a special thanks to my family. Words cannot express how grateful I am to my parents and siblings (Laila, Elias, Yasmin, Mohammed and Abdallah), for supporting and encouraging me throughout this PhD and my life in general. I am forever indebted to my parents for giving me the opportunities that have made me who I am and for their continuous and unparalleled love, help and support. In particular, I would like to express special thanks to my father who inspired me to do this PhD in the first place. I would also like to thank my grandparents for their kind support and always being there for me.



## **Declaration**

Whilst registered as a candidate for the above degree, I have not been registered for any other research award. The results and conclusions embodied in this thesis are the work of the named candidate and have not been submitted for any other academic award.

Word Count : 46839

Nabeela Altrabsheh

February 2016





# Table of contents

<b>List of tables</b>	<b>xiii</b>
<b>List of figures</b>	<b>xvii</b>
<b>Publications</b>	<b>xix</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Background and Motivation . . . . .	1
1.2 Research Importance and Questions . . . . .	3
1.3 Thesis structure . . . . .	4
1.4 Research Contribution . . . . .	5
<b>2 Previous work</b>	<b>7</b>
2.1 Students' feedback . . . . .	7
2.2 Sentiment analysis . . . . .	11
2.2.1 General vs domain-specific sentiment analysis . . . . .	12
2.2.2 Sentiment analysis applications . . . . .	12
2.3 Sentiment Analysis in Education . . . . .	13
2.4 Gaps in Previous Research . . . . .	16
2.5 Summary . . . . .	17
<b>3 Research methodology</b>	<b>19</b>
3.1 Problem Statement . . . . .	19
3.2 Aims and Objectives . . . . .	19
3.3 Research methodology . . . . .	20
3.3.1 Sentiment analysis models methodology . . . . .	20
3.3.2 Visualisation methodology . . . . .	26
3.3.3 System Evaluation Methodology . . . . .	26
3.4 Confidentiality and Ethical considerations . . . . .	27

3.5	Summary . . . . .	28
<b>4</b>	<b>Detecting polarity from students' feedback</b>	<b>29</b>
4.1	Polarity detection . . . . .	29
4.1.1	Preprocessing . . . . .	29
4.1.2	Features . . . . .	32
4.1.3	Machine Learning Techniques . . . . .	34
4.1.4	Neutral Class . . . . .	37
4.1.5	Dealing with unbalanced datasets . . . . .	38
4.1.6	Evaluation metrics . . . . .	38
4.2	Exploring N-grams and POS-Tagging features . . . . .	38
4.2.1	Data Corpus . . . . .	39
4.2.2	Sentiment Analysis Model for Students' Feedback: N-gram and POS-tagging features . . . . .	40
4.2.3	Results . . . . .	43
4.2.4	Discussion . . . . .	54
4.3	Exploring Other Features . . . . .	54
4.3.1	Data Corpus . . . . .	55
4.3.2	Sentiment Analysis Models for Students' Feedback: Other features	56
4.3.3	Results . . . . .	61
4.3.4	Discussion . . . . .	68
4.4	Optimal model for Polarity detection . . . . .	69
4.5	General Discussion . . . . .	70
4.6	Summary and Contribution of the Chapter . . . . .	73
<b>5</b>	<b>Detecting emotions from students' feedback</b>	<b>77</b>
5.1	Emotion detection . . . . .	77
5.1.1	Preprocessing . . . . .	81
5.1.2	Features . . . . .	82
5.1.3	Machine Learning Techniques . . . . .	83
5.1.4	Learning Emotions . . . . .	83
5.2	Emotion Detection Models for Students' Feedback: N-gram features . . . .	83
5.2.1	Data Corpus . . . . .	85
5.2.2	Experiments . . . . .	85
5.2.3	Results and Discussion . . . . .	86
5.3	Emotion Detection Models for Students' Feedback: Other features . . . . .	90
5.3.1	Data Corpus . . . . .	91

5.3.2	Experiments . . . . .	91
5.3.3	Results and Discussion . . . . .	95
5.3.4	Discussion . . . . .	101
5.4	Optimal Model for Detecting Students' Emotion . . . . .	103
5.5	General Discussion . . . . .	103
5.6	Summary and Contribution of the Chapter . . . . .	105
<b>6</b>	<b>Detecting sarcasm from students' feedback</b>	<b>107</b>
6.1	Sarcasm Detection . . . . .	107
6.2	Data Corpus . . . . .	110
6.3	Predicting sarcasm from students' feedback . . . . .	110
6.4	Results and Discussion . . . . .	112
6.5	Summary and Contribution of the Chapter . . . . .	115
<b>7</b>	<b>Visualisation</b>	<b>117</b>
7.1	Previous research . . . . .	117
7.2	Experiment . . . . .	120
7.2.1	Participants . . . . .	121
7.2.2	Visualisation choices . . . . .	121
7.2.3	Procedure . . . . .	124
7.2.4	Measurements . . . . .	124
7.3	Results . . . . .	128
7.4	Discussion . . . . .	132
7.5	Summary and Contribution of the Chapter . . . . .	133
<b>8</b>	<b>System integration and evaluation</b>	<b>135</b>
8.1	System Architecture . . . . .	135
8.2	System evaluation experiment . . . . .	137
8.2.1	Participants . . . . .	138
8.2.2	Procedure . . . . .	138
8.2.3	Measurements . . . . .	139
8.3	Results . . . . .	140
8.3.1	Student Questionnaire Results . . . . .	140
8.3.2	Lecturer Questionnaire Results . . . . .	142
8.4	Discussion . . . . .	144
8.5	Summary and Contribution of the Chapter . . . . .	146

<b>9</b>	<b>Summary, Contribution and Future Work</b>	<b>147</b>
9.1	Summary and Contribution . . . . .	147
9.2	Limitations . . . . .	151
9.3	Reflections . . . . .	152
9.4	Future Work . . . . .	153
	<b>References</b>	<b>157</b>
	<b>Appendix A Ethical Approval</b>	<b>173</b>
	<b>Appendix B Feature Usefulness Experiment</b>	<b>177</b>
	<b>Appendix C Emotion Feature Usefulness Experiment</b>	<b>187</b>

# List of tables

1	Abbreviations and number of features . . . . .	xviii
3.1	KDDM models . . . . .	22
4.1	Negators Examples . . . . .	31
4.2	Part-Of-Speech studies . . . . .	33
4.3	Summary of studies with features . . . . .	35
4.4	Data Sources . . . . .	39
4.5	Distribution of sentiment labels in our corpus . . . . .	40
4.6	Polarity Examples from Data Corpus . . . . .	41
4.7	Preprocessing levels and techniques . . . . .	42
4.8	Abbreviations and number of features . . . . .	43
4.9	Machine learning techniques results . . . . .	43
4.10	Performance differences between preprocessing levels for n-grams (average across all n-gram models) . . . . .	45
4.11	Performance differences between preprocessing levels for POS-tagging (av- erage across all POS-tagging models) . . . . .	45
4.12	Performance differences between preprocessing levels for all features (aver- age across all models) . . . . .	46
4.13	Noticeable preprocessing level differences . . . . .	47
4.14	N-gram results . . . . .	48
4.15	Frequent unigram words . . . . .	48
4.16	Frequent bigrams . . . . .	49
4.17	Frequent trigrams . . . . .	49
4.18	POS-tagging results . . . . .	50
4.19	Best classifiers-features combinations . . . . .	50
4.20	Highest performance for each classifier when the neutral class is included .	52
4.21	T-tests for best models . . . . .	53

4.22	Examples of studies with relatively small datasets in both domain-independent (general) and domain-dependent areas. . . . .	55
4.23	Distribution of sentiment labels in our corpus . . . . .	55
4.24	Lexicon Example . . . . .	57
4.25	Lexicon count example . . . . .	58
4.26	Features in each dataset . . . . .	60
4.27	Best models with the neutral class . . . . .	61
4.28	Best models without the neutral class . . . . .	62
4.29	Unigram models with the neutral class . . . . .	62
4.30	Unigram models without the neutral class . . . . .	62
4.31	Performance differences between preprocessing levels models with neutral . . . . .	63
4.32	Performance differences between preprocessing levels for models without neutral . . . . .	64
4.33	Difference between best models with the neutral class and unigram models . . . . .	64
4.34	Difference between best models without the neutral class and unigram models . . . . .	65
4.35	Best classifier for each of the datasets with the neutral class . . . . .	66
4.36	Best classifier for each of the datasets without the neutral class . . . . .	67
4.37	Oversampling the best models with the neutral class . . . . .	67
5.1	Summary of studies with emotions . . . . .	84
5.2	Most Frequent Emotions . . . . .	85
5.3	Less Frequent Emotions . . . . .	85
5.4	Abbreviations and number of features . . . . .	86
5.5	The classes and preprocessing levels used for the models . . . . .	87
5.6	Highest accuracy for each model . . . . .	88
5.7	Highest True Positives for each model . . . . .	88
5.8	Best overall models for identification of specific emotions . . . . .	89
5.9	Feature usefulness . . . . .	93
5.10	Experiments . . . . .	94
5.11	Best models . . . . .	95
5.12	Best models (unigrams) . . . . .	96
5.13	Difference between best models in Setups A-H and unigram models . . . . .	96
5.14	Recall for 8-Class datasets . . . . .	97
5.15	Recall for 3-Class datasets . . . . .	97
5.16	Recall for Amused dataset . . . . .	98
5.17	Recall for Bored dataset . . . . .	99
5.18	Recall for Excited dataset . . . . .	99

---

5.19	Best classifier for the 8-Class dataset . . . . .	100
5.20	Best classifier for 3-Class dataset . . . . .	100
5.21	Best classifier for Amused dataset . . . . .	101
5.22	Best classifier for Bored dataset . . . . .	101
5.23	Best classifier for Excited dataset . . . . .	102
6.1	Examples of Sarcastic Tweets . . . . .	111
6.2	Best models for each machine learning technique in sarcasm prediction . .	113
6.3	Results from previous research . . . . .	113
7.1	Visualisation advantages and disadvantages . . . . .	118
7.2	Visualisation score sums . . . . .	128
7.3	Visualisation Likeness averages results . . . . .	130
7.4	Gender Visualisation Likeness Comparison . . . . .	130
8.1	Lecture Details . . . . .	138
8.2	Lecture Details . . . . .	141
8.3	One sample Wilcoxon Signed Rank Test for individual lectures . . . . .	142
8.4	Mean for same topic lectures . . . . .	143
8.5	One-Sample Wilcoxon Signed Rank test for individual lectures . . . . .	143
8.6	One-Sample Wilcoxon Signed Rank test for individual lectures . . . . .	145





# List of figures

3.1	Generic model for sentiment analysis . . . . .	23
3.2	Model development . . . . .	24
4.1	Undersampling and Oversampling methods . . . . .	38
4.2	Optimal combination of preprocessing, features and modelling technique for our application . . . . .	53
4.3	Lexicon match example . . . . .	58
4.4	Optimal models process for polarity detection of students' feedback . . . . .	70
7.1	Neri et al. [119] bubble plot . . . . .	119
7.2	Gregory et al. [69] rose plot . . . . .	119
7.3	Demographic information about the participants . . . . .	121
7.4	All Visualisations . . . . .	122
7.5	Visualisation likeness . . . . .	129
7.6	Gender and number of emotions . . . . .	131
8.1	System components . . . . .	136
8.2	System process . . . . .	137

Table 1 Abbreviations and number of features

Category	Term	Abbreviation
Features	Unigrams .....	UNI
	Bigrams .....	BI
	Trigrams .....	TRI
	Unigrams combined with bigrams .....	UNI+BI
	Unigrams combined with trigrams .....	UNI+TRI
	Bigrams combined with trigrams .....	BI+TRI
	Unigrams combined with bigrams and trigrams	UNI+BI+TRI
	Adverbs alone .....	Adv
	Verbs, Adjectives, Adverbs, and Noun ....	VAdjAdvN
	Verbs, Adjectives, Adverbs .....	VAdjAdv
	Verbs, Adjectives, and Noun .....	VAdjN
	Verbs and Adjectives .....	VAdj
Machine learning technqie	Naive Bayes .....	NB
	Multinomial Naive Bayes .....	MNB
	Complement Naive Bayes .....	CNB
	Support Vector Machine .....	SVM
	Linear Kernel .....	LIN
	Polynomial Kernel .....	POL
	Radial Basis Kernel .....	RB
	Maximum Entropy .....	ME
	Sequential Minimal Optimization .....	SMO
	Random Forest .....	RF
Other	Neutral class .....	Neu
	Without neutral class .....	W/O
	Machine Learning Technique .....	MLT
	Student Response System .....	SRS

# Publications

The following is a list of publications and presentations made by the author of this thesis.

- Peer-reviewed conference papers
  - Altrabsheh, Nabeela, Mihaela Cocea, and Sanaz Fallahkhair. Detecting sarcasm from students' feedback in Twitter, Tenth European conference on technology enhanced learning, pp. 551-555, 2015
  - Altrabsheh, Nabeela, Mihaela Cocea, and Sanaz Fallahkhair. "Predicting learning-related emotions from students' textual classroom feedback via Twitter." The 8th International Conference on Educational Data Mining, pp. 436-440, 2015
  - Altrabsheh, Nabeela, Mihaela Cocea, and Sanaz Fallahkhair. "Predicting students' emotions using machine learning techniques." The 17th International Conference on Artificial Intelligence in Education, pp. 537-540, 2015.
  - Altrabsheh, Nabeela, Mihaela Cocea, and Sanaz Fallahkhair. "Sentiment analysis: towards a tool for analysing real-time students feedback.", IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 419-423, 2014
  - Altrabsheh, Nabeela, Mihaela Cocea, and Sanaz Fallahkhair. "Learning sentiment from students' feedback for real-time interventions in classrooms", International Conference on Adaptive and Intelligent Systems, pp. 40-49, 2014
  - Altrabsheh, N., Gaber, M. and Cocea, M. "SA-E: Sentiment Analysis for Education." The 5th KES International Conference on Intelligent Decision Technologies (KES-IDT), IOS Press, pp. 353 - 362, 2013
- Poster presentations
  - "Detecting sarcasm from students real-time feedback" , Faculty of Technology Research Day Conference, University of Portsmouth, 03/06/2015;
  - "Predicting students' emotions using machine learning techniques." The 17th International Conference on Artificial Intelligence in Education, 23/06/2015;

- 
- “Predicting learning-related emotions from students’ textual classroom feedback”, School of Computing Conference, University of Portsmouth, 18/03/2015;
  - "Modelling students’ sentiment for real-time interventions in classrooms", Faculty of Technology Research Day Conference, University of Portsmouth, 03/07/2014;
  - "Sentiment analysis: a tool for analysing real-time students feedback", School of Computing Conference, University of Portsmouth, 12/03/2014;
- Oral presentations
    - "Predicting learning-related emotions from students’ textual classroom feedback via Twitter." The 8th International Conference on Educational Data Mining, 27/06/2015;
    - "A general overview of sentiment analysis", Micro Teaching workshop, University of Portsmouth, 07/05/2015;
    - "Sentiment analysis: towards a tool for analysing real-time students feedback.", IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 10/11/2014;
    - “Learning sentiment from students’ feedback for real-time interventions in classrooms”, International Conference on Adaptive and Intelligent Systems, 08/09/2014;
    - "Sentiment analysis: a tool for analysing real-time students feedback", Faculty of Technology Research Day Conference, University of Portsmouth, 03/07/2014;
    - "Sentiment analysis for education", School of Computing Research Seminar, University of Portsmouth, 26/02/2013;

# Chapter 1

## Introduction

### 1.1 Background and Motivation

Students' feedback can highlight different issues that students may have with the lecture. One example of this is when the student does not understand part of the lecture or a specific example. Another example is when the lecturers' teaching pace is too fast or too slow. Feedback is usually collected at the end of the unit, but it is more beneficial taken in real time.

Collecting feedback in real time has numerous benefits for the lecturers and their students, such as improvement in teaching [135] and understanding students' learning behaviour [24]. Moreover, students' feedback improves communication between the lecturer and the students [24], allowing the lecturer to have an overall summary of the students' opinion.

One way of collecting feedback in real time is using Student Response Systems (SRS). Clickers, mobile phones and social media are types of SRS that have been used to collect feedback in real time. Despite their usefulness in collecting real-time feedback, SRS systems can not be used to their full advantage without support for the analysis for the collected data. For example, in a study using Twitter to collect feedback, the lecturer had to read through all the students' tweets sequentially from the beginning to understand them, causing time loss [121]. Furthermore, other research showed concern that using this tool will put such an additional workload onto their lecturers as they would require additional training to effectively use the tweets as feedback [177].

To address this problem we propose the creation of a system that will automatically analyse students' feedback in real time and present the analysis results to the lecturer. The system will visualise the students' feedback in a meaningful way giving the lecturer the most important information from the feedback. The lecturers can use this information to change their style of teaching or assessment. Time will also be taken into consideration to find

effects of students' attention at different times within a lecture or across longer periods of time, such as terms or exam periods. Eventually, the system may lead to useful information that can be used to monitor students' progress such as changing classroom materials, for example removing or adding examples. In order to create such a system, we propose the use of sentiment analysis.

Sentiment analysis is an application of natural language processing, computational linguistics and text analytics that identifies and retrieves information from text. Sentiment analysis can be applied to general data, although it is more effective when applied to specific domains [160] because word meanings and sentiment may differ across domains. Some words bear a clear positive or negative sentiment, such as the words 'good' or 'bad', while the sentiment of some other words depends on the domain [186]. An example for this is the word 'early' which may reflect negative subjectivity in education as in the instance "The lecture is too early!". Then again, when describing a parcel service such as "The parcel arrived early", this is most likely a positive sentiment.

Sentiment analysis is commonly used to identify the polarity of a given text, i.e. determining whether the expressed opinion is positive, negative or neutral [1]. However, it can also be used to detect emotions. Emotions play an important role in the learning process and thus their detection is important [50, 51, 68]. For example, positive emotions can increase students' interest in learning, increase engagement in the classroom and motivate students [51]. Happy students are also generally more motivated to accomplish their goals [118]. Troussas [174] found that students learn more efficiently and perform more successfully when they feel secure, happy and excited about the subject matter. They also stated that emotions have the ability to energize students' thinking and their emotional states can interfere with learning. For instance, emotions such as anger, anxiety and sadness can have a negative impact on students and distract their learning [174].

Sentiment analysis research has grown considerably in the last decade, mainly due to the availability of rich text resources such as social networking sites, blogs and micro-blogs and product reviews [1, 12, 27, 57, 63, 64, 92, 126, 136, 181]. Sentiment analysis in the educational domain has mainly been focused on e-learning [123, 168], with little research done on classroom feedback [118]. Although e-learning and classroom education may seem similar, they differ in the types of interactions between the students and the lecturers and in the fact that the lecturer should respond to the students' feedback in real time.

Feedback from students in the classroom settings is different from distant learners due to the different situations and issues students may have. E-learning students can have issues such as lack of interaction, while classroom feedback can be about lecture (i.e. in real time); for example, the lecturers voice being too low. When training a model, words need to be

related to the purpose of the application and in our case, the model needs to be trained using classroom feedback. To the best of our knowledge, sentiment analysis and emotion detection has not been applied for analysing students' classroom feedback before.

## 1.2 Research Importance and Questions

Applying sentiment analysis to students' real-time feedback holds many possibilities. Communication between the lecturer and the student may improve and the students' feedback may be addressed promptly. The students' feedback can be observed over time and patterns may then be identified, allowing for the lecturer to make appropriate changes.

The research is important in the sentiment analysis field, as the extensive literature review did not reveal any research exploring sentiment analysis for students' real-time feedback. In addition, research in the educational domain has been focused mainly on e-learning [83, 107, 123, 156, 168, 174].

This research is also important as there is little research in predicting emotions related to learning from students' feedback. Students' emotions detection is important as previous research has highlighted that positive emotions affect positively on students' learning, e.g. increased motivation and engagement in the classroom [51, 118].

According to the literature, the visualisation of the analysis results is also important for lecturers to make sense of it [52]. Also, it may direct researchers in the field of education into visualising sentiment analysis in a meaningful way.

The system integration and evaluation is important to the field of technology enhanced learning and it could direct researchers into ways feedback can be used to enhance learning.

In this research, we aim to answer the following questions:

- What are the best models to detect polarity from students' real-time classroom feedback?
- What are the best models to detect emotion from students' real-time classroom feedback?
- What are the best methods to detect students' sarcasm from real-time classroom feedback?
- How to present the analysis results to lecturers to facilitate decision making in real time?
- Is the system useful for the lecturers and students?

## 1.3 Thesis structure

This thesis contains research related to creating a system to analyse students' classroom feedback. This research explores sentiment analysis in the educational domain in particular students' classroom feedback. Chapter 4 covers polarity detection and chapter 5 emotion detection. Chapter 6 discusses detecting sarcasm in students' feedback. The visualisation of the detection models is presented in chapter 7, followed by the system implementation in chapter 8. A brief overview of each chapter is given in the following.

Chapter 2 discusses previous research in students' feedback and in the sentiment analysis area. It presents previous methods of collecting students' feedback and their limitations. Moreover, a discussion of sentiment analysis, which is the method we used to detect polarity and emotion from feedback is presented. Sentiment analysis in different domains and in particularly the education domain is reviewed. Gaps in the literature are identified which show the importance of this research.

Chapter 3 presents the methodological considerations of this research. A detailed discussion of the methodology and steps used for this research are presented. The relation between data mining and knowledge discovery are discussed. Previous research methodologies in data mining are reviewed. This chapter demonstrates how the chosen data mining methodology is applied to construct sentiment analysis models. This chapter also presents the methodology for the visualisation section of this research. The methodology of the system evaluation is also presented.

Chapter 4 covers the research related to the polarity detection. The chapter is initiated with a review of previous literature related to polarity detection. The following content is divided into two sub-studies. The first study is the experiment with two particular features, i.e. n-grams and Part of Speech (POS) tagging. In this study different levels of preprocessing with the most common features and machine learning techniques, i.e. Naive Bayes (NB), Complement Naive Bayes (CNB), Support Vector Machine (SVM) and Maximum Entropy (ME) used to detect polarity are explored. The purpose of this study is to identify the best combination of preprocessing, features and machine learning techniques to detect polarity. The second experiment is looking to refine the findings of the first experiment by exploring other features; more specifically, lexicons, Tweet-related, general text-related features and domain-based features. The implications of using these features in real-life settings are discussed. In this study, other machine learning techniques, i.e. Multinomial Naive Bayes (MNB), Sequential Minimal Optimization (SMO) and Random Forest (RF) are explored. By taking in consideration the two experimental studies and the practical implication, the most suitable polarity model is chosen.



Chapter 5 covers the research related to emotion detection. The chapter discusses previous literature related to emotion prediction. The emotions that are related to learning are identified. This chapter consists of two studies, which are the experiments with n-grams and other features. In the first study different n-gram combinations are explored along with different models which detect multiple emotions or single emotions. The purpose of this study is to identify whether a particular combination of preprocessing, features and machine learning techniques (i.e. NB, MNB, CNB, SVM, ME, SMO and RF) is especially suited to detect students' emotions. The second experiment is looking to improve the findings of the first experiment by exploring other features, including lexicons, Tweet-related, general text-related features and domain-based features. Finally by observing the results of the two experimental studies, the most appropriate emotion models are chosen.

Chapter 6 discusses the research related to sarcasm in sentiment analysis. This chapter explains the approach we use for labelling sarcasm in students' feedback. Different models are explored with a particular combination of preprocessing, features and machine learning techniques, i.e. NB, MNB, CNB, SVM, ME, SMO and RF. The results and their implications are discussed.

Chapter 7 covers the research related to visualisations. Previous visualisation research related to sentiment analysis and the educational domain are reviewed. In this chapter, different visualisations of the polarity and emotion detection models are presented. The visualisations are evaluated by lecturers. Finally, the results of this evaluation are discussed.

Chapter 8 covers the system architecture and system evaluation. The use of the system in real lectures is illustrated with a detailed scenario. The system evaluation section presents the users' perception of the system. The system was deployed in several real-life lectures. To assess the usability and accessibility of the system, the opinions of the students and lecturers were collected and analysed.

Chapter 9 provides a summary of the research findings and concludes the thesis. Implications and future research directions are also discussed.

## **1.4 Research Contribution**

The first contribution is to explore polarity detection in students' classroom feedback. This contribution is presented in chapter 4. The main contribution to this chapter was identifying the best model to detect polarity for students' real-time classroom feedback. This includes detecting the best preprocessing level, features and machine learning techniques. This contribution is split into several contributions. The first contribution is to identify the best preprocessing level. To the best of our knowledge, our work is the first systematic

experimentation with different levels of preprocessing aiming to identify what level is the most appropriate for our data, i.e. students' feedback, in combination with particular machine learning techniques. Another contribution is to identify the features that lead to the best performance. The last contribution is to identify the best machine learning algorithm to detect polarity.

The second contribution is to explore emotion detection in students' feedback. This contribution is presented in chapter 5. The main contribution of this chapter was to identify the best model to detect emotion students' real-time classroom feedback. This contribution is split into several contributions. The first contributions is to identify the best preprocessing level, which is explored in the first experiment. Another contribution is to identify the features that lead to the best performance, which are explored across experiments 1 and 2 of the chapter. The last contribution is to identify the best machine learning algorithm to detect emotion.

The third contribution is to explore sarcasm detection in students' feedback. This is presented in chapter 6. Similar to the first and second contributions, the best sarcasm model includes identifying the best preprocessing level, features and machine learning technique.

The fourth contribution is in the data visualisation field. When the data is presented incorrectly, it may lead to confusion and the loss of important information [52]. The contribution for the visualisation section is to create a representation of the feedback that will be useful and meaningful to the lecturer. This is presented in chapter 7. Different visualisations are created and evaluated by lecturers from different disciplines and countries. The best visualisations are identified and discussed.

The integration of the system and the evaluation of it in real settings contributes to the field of Technology Enhanced Learning. This contribution is presented in chapter 8. The system is evaluated in real-life settings to investigate its usefulness and usability to the lecturers and to find if students and lecturers could adapt to this system.

# Chapter 2

## Previous work

This chapter presents the previous literature related to students' feedback and sentiment analysis. The importance of students' feedback and methods of collecting feedback in real time are discussed in section 2.1. Sentiment analysis is a method to analyse students feedback is reviewed in section 2.2 - including previous literature related to its application in different domains. Finally, a discussion of sentiment analysis in education is presented in section 2.3.

### 2.1 Students' feedback

Students' feedback assists improvements in teaching [135]. It highlights different issues that the students may have with the lecture [38]. One example of this is when a student does not understand part of the lecture or a specific example. Students' feedback can shed light on what students do and do not comprehend and what they like and dislike about the lecture - for instance, when the lecturer's teaching pace is too fast or too slow. Additionally, students' feedback can help the lecturers understand students' learning behaviour [24].

Feedback is generally collected at the end of a semester or course [23, 34]. Students are asked to highlight different issues that may have occurred in the course [11]. However, students' feedback is more beneficial when taken in real time, as it allows lecturers to address issues in a timely manner [141].

The traditional way to obtain students' feedback in real time (i.e. classroom) is to ask students questions; for instance, if they understood a section of a lecture. However, this method does not suit everyone, such as shy students with low self-confidence [157]. Another common method to collect students' feedback in real time is using Student Response Systems (SRSs).

Student Response Systems (SRSs) is a term used to describe devices that collect real time data from participants (i.e. students). Other names for SRSs include Audience Response

Systems (ARS), Personal Response Systems (PRS) or Classroom/Student Response Systems (CRS/SRSs).

SRSs were first introduced at Stanford University in the 1966 [117]. These devices were originally wired and in the 1980s were named Personal Response Systems. In the 1990s, the devices became wireless and given the name Audience Response Systems [117]. Audience Response Systems have been used in many areas including TV programmes. However, when the devices are used in education, they are usually named Classroom/Student Response Systems (CRS/SRSs).

The main purpose of a Student Response Systems is to collect feedback. In addition, some of the SRSs allow students to respond to lecturers' questions. An important advantage of SRSs is that the student remains anonymous, therefore it can address the issue of lack of self-confidence in students to ask questions or provide feedback in-front of their fellow student colleagues. SRSs can also help in improving the student-lecturer relationship [46]. Mula and Kavanagh [117] and Mantoro et al. [104] found that students liked the use of SRSs, due to the increase interaction and engagement in the lecture. Additionally, the use of SRSs increases students' attendance and has a positive impact on students' performance [104, 117].

SRSs include handheld devices such as clickers and mobile phones that have been used previously to collect students real-time feedback. A brief overview of these is given below.

### **Clickers**

Clickers are a well-known types of Student Response System and have been used extensively over the years [104, 117, 134]. Clickers are simple handheld devices which generally contain a few buttons.

In Poulis et al. [134] clickers were very simple containing only one button that was labelled 'yes', which could be used to respond to the lecturers' enquiries. The lecturer asked the students questions and then they responded by interacting with the clickers - for instance, if the lecturer wanted to know if the students understood a certain point, so that s/he can decide whether to move to the next topic. Mula and Kavanagh [117] also used clickers which were comprised of 12 buttons, allowing a wider variety of questions to be posed.

One the many advantages of clickers is that they allow students to focus longer on the material and learn it through participation, rather than focusing on taking notes throughout the entire lecture [125]. A further advantage is student engagement [46]. Denker [46] claimed that participation was higher using clickers compared with using "raising hands" method.

On the other hand, clickers cost institutions money and students may lose or break them, or forget to bring them to class [40]. Moreover, clickers can also distract students in class [46].

### Mobile Phones

Mobile phones are a type of SRSs device that are popular with students. A study by Scornavacca et al. [146] found that 98% of students own a mobile phone. Students usually have their mobiles with them in class making them an expedient device to ask them to use them for feedback. So and Wah [154] found in their study that more than 50% of participants agreed that mobile phones improves learning.

Despite their popularity, the use of mobile phones in the classroom also has some negative outcomes. For instance, mobiles that ring during class are obviously distracting and it can be tempting for students to use them for other activities such as sending messages to their friends [10, 146]. Another disadvantage is that students may flood the system with inappropriate messages [10].

There are two different approaches to use mobile phones as SRSs: Short Message Service (SMS) and clicker applications (only with smartphones). The following is a brief explanation of both.

- Short Message Service (SMS): Leong et al. [94] collected students' feedback through SMS. In their research, they created a model that collected feedback from students via SMS messages. The aim of their research was to improve the delivery of the lesson by finding out students' opinions.

They found that taking feedback via SMS had many flaws, such as the limitation of the message space to a certain number of characters. Moreover, the text contained spelling mistakes, however, this was addressed using "The Corrected Model". The Corrected Model converted words that were misspelled into similar words such as slp to sleep. The authors claimed to use sentiment analysis to analyse the SMS messages. However, they did not include any details describing the sentiment analysis process or evaluation, therefore this was not discussed further in the sentiment analysis section of this chapter.

Despite the advantages of Leong et al. [94] study, the cost of the SMS texting service was not taken into consideration. Kinsella [86] suggested that lowering the cost barrier may help to increase feedback during lectures.

- Clicker Application: Instead of using a proprietary handheld device, the clicker was presented as an application on students' mobile phones. The clicker application addresses the problem of the clickers' high costs. However, similar to the clicker devices themselves, the interaction is limited to a number of options.

One example of the clicker applications is Teevan et al. [162]. In their study “the Crowd Feedback” system, they aimed to collect students’ feedback. Students could choose to either ‘Like’ (a green thumbs up button) or ‘Dislike’ (a red thumbs down button), when the lecturer asked them specific questions. The lecturer’s phone would vibrate if something interesting was found from analysing that feedback.

The system did not only provide feedback for the lecturer but for the students as well. The system also revealed that the students gave feedback all throughout the lecture. The study acknowledges that feedback would be more beneficial if it were richer, extending beyond just simply providing ‘like’ or ‘dislike’ answers.

Another example of the use of Clicker Applications is in Mantoro [104], which consisted of an application via a web-based system. The students could use either their own mobiles or PCs to answer the lecturer’s questions. They claimed that their application has an advantage over clickers as it addresses the high cost of clickers. Their research only proposes the application and an experiment to test its effectiveness in real-life settings was not included.

One problem with installing applications on mobile phones is compatibility with the operating platform of the mobile phone due to the wide variety of systems such as Android, IOS and Windows.

## **Social Media**

Social media is one of the most used communication media in the modern era. Statistics reveal that 93.5% of 18 year olds and 95.4% of 19 year olds in the USA were found to use social networking on a regular basis [121]. The same survey revealed 52.1% of academics had used Twitter in 2010. In addition, over 470 universities worldwide use social networks such as Facebook and Twitter to communicate with their students [121].

The use of Twitter in education has the advantage that students are familiar with the tool and training would not be needed [121]. Other benefits of using Twitter as means of providing feedback is that it is free and is accessible from students’ own mobiles on the university wireless network.

Twitter has been found to have a positive impact on students and to increase their motivation [48, 70]. When students use something they are familiar with and enjoy, they will participate more in providing feedback on the lecture in real time [48].

Retelny et al. [139] used Twitter in the context of lectures and students found these lectures engaging and interactive. Their study found that using Twitter encouraged student participation and that students felt Twitter helped them grasp learning concepts. Dhir et

al. [48] highlighted some advantages of Twitter that include: students bonding with the lecturer, an improved student-lecturer relationship, students improving their writing skills, students solving problems in a timely manner and more privacy for students asking questions.

Twitter is capped to only 140 characters; this can be both an advantage and disadvantage. It is an advantage because students tend to write only the important details and keep the tweet short. However, cutting sentences short may mean the true meaning of the sentence is not communicated.

Twitter has some disadvantages such as it being a potential distraction for the students or the lecturer due to multitasking [48, 59]. The use of Twitter also raises concerns such as privacy, time-wasting and even addiction [70]. Because Twitter is an all-purpose tool, students can use it for other purposes such as messaging their fellow students. One way for limiting this is to set times for students to tweet their feedback.

In Novak and Cowling [121], Twitter was used as a tool to collect feedback in the classroom. Despite having many advantages such as class interaction and engagement, there were some difficulties. For example, the lecturer had to read through all the students' tweets sequentially from the beginning to understand them, causing time loss. Furthermore, some of the campuses involved in Novak and Cowling [121] study claimed that using Twitter would put such an additional workload on their lecturers that they would require additional training to effectively use the tweets to provide feedback.

There are not many methods used to automatically analyse text data. One of the most well-known fields examining text analysis is Natural language processing. Natural language processing consists of various methods to analyse text, such as pattern and trend finding using word frequencies or information retrieval and extraction. One popular method is sentiment analysis, which is a simple method classifying text to polarity (positive, negative or neutral). This method is considered the most suitable method due to its similarity to the clicker application, employing 'like' or 'dislike' buttons and making it possible to extract useful information from the text. Sentiment Analysis is further explained in the following section.

## 2.2 Sentiment analysis

Sentiment analysis is an application of natural language processing, computational linguistics and text analytics that identifies and retrieves sentiment polarity from text by studying the opinion [115]. One task in sentiment analysis is classifying the polarity of a given text, i.e. determining whether the expressed opinion is positive, negative or neutral [1].

Sentiment Analysis can also be used to extract more specific emotions. These can be general, such as joy, anger, sadness, disgust, fear and surprise [132] or specific to the domain, for instance learning emotions such as frustration, boredom and confusion [50].

### **2.2.1 General vs domain-specific sentiment analysis**

Some studies have applied sentiment analysis in general without specifying the domain [1, 12, 63, 64, 92, 126, 136, 181]. However, previous research has shown that sentiment analysis is more effective when applied to specific domains [63, 177]. Some word meanings and their sentiments may be the same across different domains, as in the case of the words ‘good’ and ‘bad’, however the sentiment of some words may vary depending on the domain [186]. For example, the word ‘revision’ may reflect negative polarity in the educational domain; for instance: “Big exam tomorrow, so much revision to do”. On the other hand, in the law domain, the word ‘revision’ may reflect positive polarity; for example: “The law is currently undergoing revision to close a rather strange loophole”. Many applications are domain-specific, as outlined in the following subsection.

### **2.2.2 Sentiment analysis applications**

Sentiment analysis can be applied for different purposes such as political campaigns and riots, however, it is most often applied on reviews. The reviews can be from different domains including movies [9, 131, 185], advertisements [76], products [27], cars [57], smart phones [27], tourism [32] and e-learning [123, 168]. We focus on outlining previous work related to sentiment analysis on Twitter data.

#### **Sentiment analysis on movie reviews**

Asur et al. [9] applied sentiment analysis on Twitter to discover feedback on movies in advance of their release. This study predicted box-office revenues of forthcoming movies and found a strong correlation between the amount of attention a given topic has in Twitter (i.e. forthcoming movie) and its ranking in the future.

Pang and Lee [131] also performed sentiment analysis on movie reviews using machine learning techniques. They found that using machine learning techniques to classify data outperforms human classification.



**Sentiment analysis on advertisement reviews and sports**

Hill et al. [76] explored real-time advertisements at a baseball game. They found that the number of tweets made by the audience about advertisements was correlated with the ultimate success of their underlying products.

Zhao et al. [190] performed sentiment analysis on Twitter to find peoples' opinions to major events in live broadcast sports games in real time. Their research indicated that Twitter might also be a good tool to extract opinions about TV programs.

**Sentiment analysis on products including cars and smart phones**

Chamlertwat [27] collected tweets related to smart phones, finding that sentiment analysis on Twitter could detect customers' opinions towards the product features such as apps, screen and camera. This could allow consumers to gather information regarding products without disturbing other consumers by asking them directly for reviews.

Gamon et al. [57] applied sentiment analysis on car reviews, revealing that sentiment analysis with an appropriate visualisation provides a powerful tool for exploring customer feedback. This research points out the importance of visualisation in understanding sentiment analysis results.

**Sentiment analysis on tourism**

Claster et al. [32] found that sentiment analysis could be used to extract users' opinions about tourism and local attractions to better advise the officials of that country.

Kasper and Vela [82] applied sentiment analysis to hotel reviews. Their research allowed consumers to have a summarized opinion of the hotel before booking including the positive and negative aspects of the hotel. Additionally, it enables hotel managers to follow and detect issues to improve their ratings.

## **2.3 Sentiment Analysis in Education**

In the educational domain, most research has been on sentiment analysis in e-learning [83, 107, 123, 156, 168, 174], with only one study carried out on classroom feedback [118]. These studies are reviewed in this section.

Ortigosa et al. [123] detected polarity (i.e. positive, negative and neutral) from Facebook in the context of e-learning. The researchers also detected emotions that were categorised into two classes: positive and negative. The positive class consisted of positive emotions, positive sentiment and 'optimistic' emotion. The negative class included negative emotions, 'anger',

‘sadness’, ‘death’ and ‘swearwords’. Their research examined Spanish data collecting 1000 instances of positive, 1000 instances of negative and 1000 instances of neutral from Facebook text messages and statuses. For preprocessing (preparing the data), they converted text into lower case, identified repeated letters (through regular expressions) and emoticons and assigned a score to them for polarity strength and then used tokenization. Two methods were used to detect sentiment; a lexicon-approach and a combined-approach (i.e. lexicon-based and machine learning-based). For the lexicon-approach, they tokenized the data and matched it to a lexicon (a dictionary of items and their polarity score) summing the scores to detect polarities. As for the combined-approach, word vectors were used as well as matches from the lexicon for features (inputs of the classifier) and Java implementation of the C4.5 algorithm (J48), Naive Bayes (NB) and Support Vector Machine (SVM) were used as machine learning techniques (classifiers). Their results indicated that using a combined-approach led to a better performance than using a lexicon-approach alone. Results were evaluated using accuracy and SVM led to the highest accuracy at 83%. Ortigosa et al. [123] pointed out that detecting polarity could be used in adaptive e-learning systems to support personalized learning. They also pointed out that this can help lecturers in e-learning courses, where face-to-face contact is less frequent.

Similarly, Martin et al. [107] applied sentiment analysis to writings in Facebook walls in the context of e-learning to detect polarity and classified each sentence as positive, neutral or negative. Their research also used data in Spanish language. Three thousand instances of positive, neutral and negative instances were collected. Emotions were detected that were categorised similar to Ortigosa et al. [123] study (i.e. Positive: positive emotions, positive sentiment, optimistic and Negative: negative emotions, anger, sadness, death and swearwords). Tokenization was used and text was converted to lower case to preprocess the data. Several features were used which included: lexicon-based features, negation, POS (Part-Of-Speech)-tagging (i.e. noun, adjective, interjection or verb), emoticons and spelling check. To detect sentiment they used only a lexicon-based method that included summing the words found in the lexicon polarity scores. Accuracy was used to evaluate their models and consequently the predicted accuracies of the positive, negative and neutral classes were 96%, 80% and 83% respectively.

Troussas et al. [174] also explored students’ polarity through Facebook in the context of a language learning application. They used a machine learning approach to detect the polarity (i.e. positive and negative classes), which included Naive Bayes, Rocchio and Perceptron classifiers. The data was divided as 50% training and 50% testing and unigrams (single words) were used as features. Their models were evaluated by precision, recall and F-score that are standard Data mining evaluation techniques explained more in chapter 3. As a result,

the study found that using Naive Bayes and Rocchio classifiers performed the best giving F-scores of 72% and 74% respectively. Although their research was focused on polarity, they also reported on the importance of detecting students' emotions and its impact on their learning and actions. Troussas et al. [174] argues that students who are working to cope with emotions may not have sufficient resources available to engage in learning. Moreover, the authors argue that emotions such as anger, anxiety and sadness have a negative influence on students' learning and can distract students. On the other hand, students that are happy usually perform better [174]. However, students who are overly excited or enthusiastic may work carelessly or quickly instead of working methodically or carefully [174]. Detecting these emotions can help lecturers take action by focusing attention on certain students.

Kechaou et al. [83] outlined the importance of detecting polarity from e-learners' feedback from e-learning courses, with most of the communication between lecturers and students being through text form such as emails and forums. The research focused on analysing sentiments for the purpose of developing and improving of e-learning systems. Data was classified into positive and negative classes. Two thousand instances of positive and negative data from e-learning blogs and text reviews from Moodle forums were collected. A machine learning approach was employed that included Hidden Model Markov (HMM) and Support Vector Machine classifiers. The data was split into 80% training and 20% testing. To select the most relevant features, they used Information Gain, Chi Square and Mutual Information tests. The sentiment analysis models were evaluated by precision, recall and F-score and a F-score rate of 80% was achieved.

Song et al. [156] also explored sentiment analysis on e-learning systems to find students' opinions about courses, lecturers and e-learning systems. This research could benefit both lecturers and e-learning software developers. They detected polarity (i.e. positive and negative) from the sentiments. This study was conducted on the Chinese language data, with a Chinese corpus consisting of 7,324 sentences from 446 review articles gathered from the learning domain. Negation and POS-tagging (i.e. degree adverbs) were used as features. Similar to Kechaou et al. [83], a machine learning approach was utilised, including only the SVM classifier. The models were evaluated using precision, recall and F-score and they achieved a F-score result of approximately 90%.

Tian et al. [168] detected e-learners' emotions (i.e. Love, Joy, Anger, Frustration) from their texts in a chat system. They highlighted the benefits of detecting emotions through text, which is a less expensive method compared to intelligent tutoring systems that require expensive equipment and technologies. The authors collected 1,455 sentences and labelled their instances using ChineseParser, a software they developed. POS-tags were used as features and C4.5 decision tree as a classifier. Weka was used to perform the experiments

and the models were evaluated using precision, recall and F-score. Tian et al. [168] found that the ‘anger’ emotion could be detected more easily than others and resulted in the highest F-score at 95%.

There was only one classroom-related study conducted by Munezero et al. [118] that detected students’ emotions from their learning diaries. The lecturer uploaded the students’ diaries into the system on three occasions over different periods. Eight emotions were detected: joy, sadness, fear, anger, anticipation, surprise, disgust and trust. Their research was implemented using a lexicon-based approach. Moreover, a general lexicon (i.e. NRC word-emotion association lexicon) was used. For each of the emotions, the polarity scores of the words found in the lexicon were averaged and the sentence was labelled according to the majority score. They visualised the sentiment analysis results using word cloud, radar, bar and line charts. The visualised data was further presented over periods of time (i.e. monthly). One limitation to Munezero et al. [118] study was that it did not include any evaluation of the sentiment analysis models or the visualisations.

## 2.4 Gaps in Previous Research

Sentiment analysis models perform better when trained for the domain in which they are used [63, 160, 177]. Therefore, they need to be trained with data that is related to the purpose of the application. After reviewing previous works related to the educational domain, some limitations were found. Some of the studies such as Troussas et al. [174] and Kechaou et al. [83] used percentage splitting to evaluate their models. Cross validation is recommended in the data mining models instead of percentage partitioning as it is a non-biased approach which selects data randomly into folds and tests the models across all the data.

In addition, many of these works such as Martin et al. [107] and Munezero et al. [118] use a lexicon-based approach that is less accurate than using machine learning [123]. In comparison to Munezero et al. [118] research which was the only one that was classroom based, the research presented in this thesis focuses on real-time feedback during the lecture by allowing students to give their own feedback via Twitter using their mobiles, tablets or laptops. This could be a more efficient method than lecturers uploading the feedback, which could cause time-loss.

Munezero et al. [118] used a general lexicon (i.e. not domain specific) which could lead to inaccurate classification of words given the different word meanings across domains. Moreover, their study was not evaluated using common data mining and sentiment analysis evaluation metrics (i.e. accuracy, precision, recall and F-score), and they did not include any details of evaluation or testing the performance of the models. Munezero et al. [118] did

visualise the results for sentiment analysis, however, they did not provide any information or study to evaluate the visualisation usefulness or usability.

To the best of our knowledge, there is no model fit for analysing students' classroom feedback. In addition, there is a need to study the best method to visualise sentiment analysis results to the lecturer. Both of these aspects are addressed in this thesis.

## 2.5 Summary

In this chapter, the advantages of students' real-time feedback and previous collection methods have been reviewed. Previous research suggests that analysing feedback in real time is both stressful and time-consuming and that sentiment analysis can be used to address this.

The reviewed literature related to sentiment analysis includes the application of sentiment analysis in general or to specific applications. The disadvantages of applying sentiment analysis in general were presented in section 2.2.1. Additionally, a detailed description of sentiment analysis in the educational domain was included in section 2.3.

Most of the research on sentiment analysis in the educational area has been focused on the e-learning area. However, there are differences between e-learning and classroom-based sentiment analysis given the words used differently in each area. Only one study was related to classroom-based sentiment analysis where the authors applied sentiment analysis to students' diaries. Limitations from previous sentiment analysis research in the educational domain were reviewed and the gaps addressed by the research thesis were outlined.

The literature reveals that there is no model or research on sentiment analysis for students' real-time classroom feedback that uses machine learning techniques. There is also need to explore and create visualisations to present the results to lectures in a way that can be useful to them.



# **Chapter 3**

## **Research methodology**

This chapter presents a detailed description of the research methodology for this thesis. In section 3.3, the details of the methodological steps for this research are presented. Before presenting details of the methodology, the problem statement, research aims and objectives are highlighted in the sections 3.1 and 3.2.

### **3.1 Problem Statement**

The previous literature suggests that analysing feedback manually is both time-consuming and stressful. In this thesis, this issue is addressed by proposing a system that can automatically analyse students' feedback and present it to the lecturer in real time. To create the system, sentiment analysis is used to analyse the feedback. The previous literature does not reveal any sentiment analysis models suitable for analysing students' feedback. Therefore, this research explores different attributes of sentiment analysis models to identify new models. There is also lack of research in visualising this kind of data, therefore, in this research different ways to present the data in a meaningful way are examined.

### **3.2 Aims and Objectives**

This PhD thesis focuses on creation of a system that analysis students' feedback and presents it to the lecturer in real time. The system consists of identifying sentiment analysis models that give optimal results, and creating visualisations that are both meaningful and useful for the lecturer. To meet the above mentioned aims, the following objectives were defined:

1. Identify the best suited models to predict polarity and emotions, which includes identifying the best combination of preprocessing techniques, features and machine learning techniques.
2. Investigate different ways of representing the results and create visualisations that facilitate the understanding of the sentiment analysis models results.
3. Evaluate the system in real lectures.

### **3.3 Research methodology**

This section presents a detailed description of the methodology for each of the three objectives outlined above.

#### **3.3.1 Sentiment analysis models methodology**

There are different approaches to create sentiment analysis models: lexicon based-approach, machine learning approach and hybrid approach.

Lexicon based-approach is a simple approach where the words in a new sentence are searched for in a lexicon. The lexicon contains words with their polarity label or a number reflecting how much the word expresses each polarity/emotion. The polarity/emotion label can be assigned to the sentence according to the majority score. For example, if the majority of the words were positive or the positive total was higher than the negative total, the sentence will be labelled as positive. This approach is not that accurate and can be misleading due to its dependency on the lexicon. If the words were not found in the lexicon, the sentence will remain unlabelled.

The machine learning approach is a more efficient method to detect polarity/emotion. This approach uses machine learning techniques which are classifiers to predict the polarity/emotion. This approach is more accurate than the lexical-based approach [123].

The hybrid-approach or combined-approach uses the lexicon-based and machine learning approaches together. This is usually implemented by including the values/results found from the lexicon as inputs (i.e. features) to the machine learning classifier. For this research, the machine learning and hybrid-approaches are utilised. Machine learning is a method of Data Mining (DM).

Data Mining aims to explore and analyse datasets to extract patterns and knowledge and find relations between attributes [53]. DM is also known as knowledge extraction, information discovery, information harvesting, data archaeology and data pattern processing [93].



Knowledge Discovery (KD) is a process of finding new knowledge about an application domain. DM is one of knowledge discovery's steps which applies a certain discovery task in an application domain.

Knowledge Discovery and Data Mining (KDDM) is the process of applying KD to any data source. KDDM includes the entire knowledge extraction process, which includes developing efficient algorithms to analyse the data, interpreting and visualising the results, and modelling the interaction between human and machine [93].

We can apply KDDM models steps to construct sentiment analysis models. There are many existing KDDM models such as Fayyad et al. [53], Cabena et al. [22], Anand and Buchner [5], CRISP-DM [152], Cios et al. [31], and the Generic model [93]. The steps of these models are found in Table 3.1. The common steps through the five models are: Domain Understanding, Data Preparation, DM and evaluation of the Data Knowledge. Fayyad's nine-step model is different from the other models as it performs activities related to the choice of the DM step late in the process. All other models perform this step before preprocessing the data. This approach is better as when the data is correctly prepared for the DM step, there is no need to repeat any other process steps [93]. One of the benefits of Fayyad's model is that in case prepared data is not suitable for the tool of choice, it includes looping back to the second, third or fourth steps [93]. This is also beneficial for data that requires extensive preprocessing.

Cabena's, Cios and the CRISP-DM model are very similar, however Cabena's omits the Data Understanding step, which is a necessary step according to Kurgan and Musilek [93]. Cabena's model is suitable for applications where data is virtually ready for mining before the project starts.

Anand and Buchner's model is very detailed and includes steps of the early phases of the KDDM process. However, it does not include necessary activities for putting the discovered knowledge to work.

The generic model was introduced by Kurgan and Musilek [93]. It differed from all the other models as they all involved complex and time-consuming data preparation tasks. The processes in other models include iterations and loops between most of the steps. The generic model consists of six steps and is similar to Cios et al. [31] and CRISP-DM models. It was built based on older models and its six steps can fit all other models with a minor modification that it combines several original steps into a major step. Thus, the reason for choosing it in this research. As shown from Figure 3.1, when utilising the generic model process to the identification of sentiment analysis models, the following steps are applied:

1. Understanding the educational domain and the system (application) as a whole: this step is important and determines the other steps

Table 3.1 KDDM models

Fayyad et al. [53]	<ol style="list-style-type: none"> <li>1. Developing and Understanding of the Application Domain</li> <li>2. Creating a Target Data Set</li> <li>3. Data Cleaning and Preprocessing</li> <li>4. Data Reduction and Projection</li> <li>5. Choosing the DM Task</li> <li>6. Choosing the DM Algorithm</li> <li>7. DM</li> <li>8. Interpreting Mined Patterns</li> <li>9. Consolidating Discovered Knowledge</li> </ol>
Cabena et al. [22]	<ol style="list-style-type: none"> <li>1. Data Preparation</li> <li>2. DM</li> <li>3. Domain Knowledge Elicitation</li> <li>4. Assimilation of Knowledge</li> </ol>
Anand and Buchner [5]	<ol style="list-style-type: none"> <li>1. Human Resource Identification</li> <li>2. Problem Specification Data Prospecting</li> <li>3. Domain Knowledge Elicitation</li> <li>4. Methodology Identification</li> <li>5. Data Preprocessing</li> <li>6. Pattern Discovery</li> <li>7. Knowledge Post-processing</li> </ol>
CRISP-DM [152]	<ol style="list-style-type: none"> <li>1. Human Resource Identification</li> <li>2. Problem Specification</li> <li>3. Data Prospecting</li> <li>4. Domain Knowledge Elicitation</li> <li>5. Methodology Identification</li> <li>6. Data Preprocessing</li> <li>7. Pattern Discovery</li> <li>8. Knowledge Post-processing</li> </ol>
Cios et al. [31]	<ol style="list-style-type: none"> <li>1. Understanding the Problem Domain</li> <li>2. Understanding the Data</li> <li>3. Preparation of the Data</li> <li>4. DM</li> <li>5. Evaluation of the Discovered Knowledge</li> <li>6. Using the Discovered Knowledge</li> </ol>
Generic model [93]	<ol style="list-style-type: none"> <li>1. Application Domain Understanding</li> <li>2. Data Understanding</li> <li>3. Data Preparation and Identification of DM Technology</li> <li>4. DM</li> <li>5. Evaluation</li> <li>6. Knowledge Consolidation and Deployment</li> </ol>

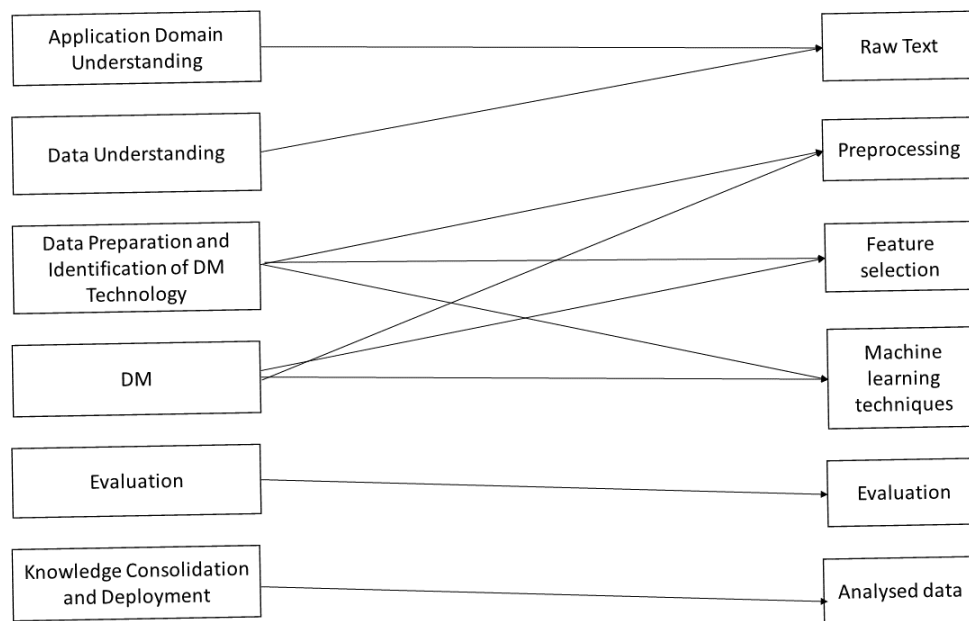


Fig. 3.1 Generic model for sentiment analysis

2. Understanding the students feedback: this step allows us to examine the data and plan for preprocessing
3. Data preprocessing, feature selection, and determining the machine learning techniques: this step is to explore which of preprocessing, feature selection, and determining the best performing machine learning techniques for the application (i.e. leading to the highest accuracy)
4. Sentiment analysis step (DM step): this step includes testing sentiment analysis models on the data
5. Evaluation: evaluating the models using the data mining evaluation metrics, which are accuracy, precision, recall and F-score
6. Knowledge Consolidation and Deployment: this step allows us to extract knowledge from the proposed system such as labelling the students' feedback.

In other words to construct sentiment analysis models using a machine learning approach, the domain and data need to be understood then four main steps are applied: preprocessing the data, selecting the features, applying the machine learning techniques and evaluating the results. These steps involved in model development are illustrated in Figure 3.2. An overview of previous sentiment analysis research related to these steps is given in the following subsections.

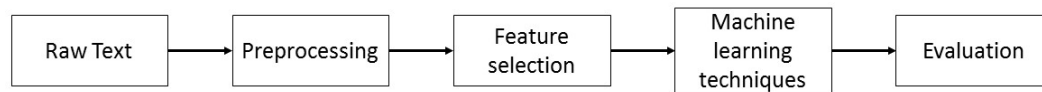


Fig. 3.2 Model development

### ***Preprocessing***

Preprocessing is the process of cleaning the data from unwanted elements. It increases the accuracy of the results by reducing errors in the data [3]. Not using preprocessing, such as spelling corrections, may lead the system to ignore important words. On the other hand, overusing preprocessing techniques may sometimes cause loss of important data. One example illustrating this is removing punctuation in “what does this example mean???”; the question mark (and its repetition) may indicate confusion.

There are many general preprocessing techniques, of which the most common are: tokenization, covert text to lower or upper case, remove punctuation, remove numbers, remove repeated letters, remove stop words, stemming and negation [27, 32, 115, 126, 136]. Some of these preprocessing techniques are found in previous sentiment analysis research, e.g. Mouthami et al. [115], Pak and Paroubek [126, 127], Chamlerwat et al. [27], Claster et al. [32] and Prasad [136].

Preprocessing data from social media is different due to the special symbols existing in social media data such as emoticons, hashtags and chat language. Social media is a popular tool to collect peoples’ opinions as it provides the users a simple interface, enabling and encouraging them to express their sentiments. Twitter is one of the largest and most popular social media platforms for capturing opinions [127]. Analysing Twitter data is different from analysing basic text.

Some of the most common preprocessing techniques that are used for Twitter data are removing hashtags [92], removing URLs [92, 126, 136, 190], removing retweets [190], identifying emoticons [1, 92, 136, 190], removing user mentions in tweets [126, 190], removing Twitter special characters [92] and handling slang/chat language [1]. These techniques are described in more details in chapters 4 and 5.

### ***Features***

Features give a more accurate analysis of the sentiments and detailed summarisation of the results [186]. The most common features are n-grams [1, 63, 180] and POS (part-of-speech) tagging [131, 181].

Other features that are less common can include: lexicon-based features, Twitter-related features (i.e. number of hashtags and emoticons) and punctuation-related features (i.e. number of punctuations and particularly question and exclamation marks). More details about these features are presented in the experiments in the chapters 4 and 5.

### ***Machine Learning Techniques***

The most popular machine learning techniques for sentiment analysis are Naive Bayes (NB) [17, 57, 63, 111, 129, 160], Multinomial Naive Bayes (MNB) [63, 170], Support Vector Machines (SVM) [17, 44, 63, 111] and Maximum Entropy (ME) [63, 111].

### ***Evaluation***

The most common evaluation metrics used in sentiment analysis are accuracy, precision, recall and F-score, as outlined below. All the definitions relate to the following concepts, explained below by taking as an example a classification into two classes, i.e. positive and negative.

1. True positives (TP) or true positive rate represents the number of correctly classified instances for the positive class divided by the total number of instances, i.e. the instances classified as positive that are in reality positive;
2. False positives (FP) represent the number of incorrectly classified instances for the positive class divided by the total number of instances, i.e. the instances classified as positive that are in reality negative;
3. True negatives (TN) represent the number of correctly classified instances for the negative class divided by the total number of instances, i.e. the instances classified as negative that are in reality negative;
4. False negatives (FN) represent the number of incorrectly classified instances for the negative class divided by the total number of instances, i.e. the instances classified as negative that are in reality positive;

The accuracy metric indicates how many instances were correctly classified across all classes:  $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$ . Precision represents the fraction of retrieved instances that are relevant, e.g. how many of the instances classified as positive are actually positive:  $Precision = \frac{TP}{TP+FP}$ . Recall is the fraction of relevant instances that are retrieved, e.g. how many of the instances that are actually positive were classified as positive:  $Recall = \frac{TP}{TP+FN}$ . The F score is a metric of accuracy that combines precision and recall:  $F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$ . Precision, recall and F scores for a classifier are obtained as the average of these values for all the classes.

### 3.3.2 Visualisation methodology

Visualising the data involves presenting it in a meaningful way to the user. Knight [87] claims that the visualisation is considered effective if users can achieve results using it, without extra cost. The visualisation should be easy to understand and should serve its purpose without any external help or information. Experimentation with different visualisation techniques is needed, as various data requires different representations for easy comprehension. This research applies a critical comparison of different known visualisations used in sentiment analysis, for instance bar, line, pie and scatter charts. Additionally, new ways to represent the data are explored. Multiple visualisations that are familiar to the users (i.e. lecturers) and which originate from previous visualisations are created. This provides lecturers with different options, as some favour different visualisation representations over others.

To evaluate the visualisations, a participatory design was chosen, which is a qualitative research method allowing lecturers to participate in the design by providing feedback and suggestions for new visualisations and ensuring that the visualisations meet their needs and are usable. Interviews and questionnaires were used for this study. The interview process started with the purpose of the study and the lecturers were presented with several visualisations. The lecturers then filled out a questionnaire and were asked if there were any suggestions for changes in the current visualisations or any new ideas for visualisations that may have been overlooked. Data was collected from lecturers from different countries. The lecturers that were not present locally were not interviewed. Instead they were provided with the same information given in each interview in writing along with the questionnaire. Garrote et al. [58] also used questionnaires in interviews in a similar context to find lecturers' opinions about the use of learning management systems. Interviews allow lecturers to elaborate more about their answers and the evaluation is not limited to the questionnaire questions. Questionnaires allows us to quantify the results and to find the majority opinions about the visualisation designs.

To ensure the robustness of the results, lecturers from different countries and disciplines were selected. The visualisation methodology is explained in more detail in chapter 7.

### 3.3.3 System Evaluation Methodology

To evaluate the full system, combining the sentiment analysis models with the visualisation of results, real-life settings were used. The inputs of the system include the students' feedback and the output of the system are visualisations of the polarity and emotion detection results. To evaluate the system, usability, usefulness and adaptability of the system are assessed.

The usability of the system is tested to assess how users perceive and use it. This includes the interaction of the lecturers with the system, and if they are able to use and integrate it in their lectures. It also includes exploring how many students provide feedback in the lecture.

Usefulness is assessed to measure how useful the tool will be to lecturers. This includes how useful they find the polarity and emotion prediction models and the visualisations in general.

The adaptability testing is used to investigate if the lecturers/students adapt to the system. This is to explore if lecturers can integrate this system in their everyday lectures, and if they will use the system again. This is also to examine students' opinions about providing feedback and their willingness to do so in future lectures.

Questionnaires were used to assess all the aspects mentioned above, due to it being easy to quantify results and can be distributed among a larger number of people. Questionnaires are a quantitative method and have previously been used to assess educational systems. Liaw [98] used questionnaires to test students' satisfaction in blackboards. Likewise, Harrington et al. [74] used questionnaires to see students' opinions in a Web-Based Course Management system. Garrote et al. [58] used questionnaires in interviews to judge lecturers' attitudes about the use of learning management systems.

The questionnaires were created by observing previous system evaluation questionnaires, for instance "Usefulness, Satisfaction and Ease of use" questionnaire created by Lund [103] and "Computer System Usability" questionnaire from Lewis [96]. A Likert scale was used in the majority of the questions, similar to most of the previous questionnaires [96, 98, 103]. In Likert scale questions, respondents specify their level of agreement or disagreement on a scale capturing the intensity of their opinion.

Two questionnaires were distributed, one for the lecturers (the users of the system) which is an in-depth questionnaire. The second was for students, which is a short questionnaire exploring whether they participated in providing feedback and if they adapted to the new teaching style (i.e. providing feedback and immediate action/changes to the lecture). More about the evaluation of the system will be presented in chapter 8.

### **3.4 Confidentiality and Ethical considerations**

Due to the research involving student feedback, there needs to be confidentiality and ethical considerations. First the students feedback stays confidential and is not shared with anyone or published. The data is not accessible to any person other than the researcher. The reason for collecting feedback is explained to the students beforehand and the participation is voluntary. The students remain anonymous as a Twitter account is set for all of them, therefore they do

not use their personal accounts unless they choose to. Either way, when storing the data, the name is discarded and only the feedback is taken. The researcher received ethical approval for this research through the University of Portsmouth procedure - please see the ethical certificate in Appendix A.

### **3.5 Summary**

This chapter reviews research methodology for this thesis. The overall methodological approach was presented with other methodological details that were specific to the different research objectives. Moreover, the confidentiality and ethical considerations are outlined.



# **Chapter 4**

## **Detecting polarity from students' feedback**

This chapter presents two studies for polarity detection from students' feedback. Previous literature related to the sentiment analysis steps in polarity detection are reviewed. This chapter presents two experiments: the investigation of n-gram and POS-tagging and the investigation of other features. This chapter is concluded with the best model found to detect polarity from students' real-time feedback. Finally, a summary of the chapter and contribution are presented.

### **4.1 Polarity detection**

The aim of this research is to identify a sentiment analysis model that is tailored for the educational domain, particularly for students in-class real-time feedback. There are four main steps to identify sentiment analysis models using the machine learning approach for polarity detection: preprocessing the data, selecting the features, applying the machine learning techniques and evaluating the results. Previous research related to these are discussed in the following subsections.

#### **4.1.1 Preprocessing**

Preprocessing is an important step to prepare the data for analysis. There are many general preprocessing techniques in polarity detection, of which the most common are: tokenization, converting text to lower or upper case, removing punctuation, removing numbers, removing repeated letters, removing stop words, stemming, and negation [27, 32, 115, 126, 136]. Some

of these preprocessing techniques are found in previous sentiment analysis research, as outlined below:

1. Tokenization: Breaking a sentence into words, phrases, and characters. Tokenization was used in Mouthami et al. [115]; Pak and Paroubek [126]; and Chamlerwat et al. [27].
2. Covert text to lower or upper case: Converting the letters into upper or lower case to match them with other occurrences in the training data. A word in capitals sometimes suggests strong emotion [136, 190]. One example illustrating this is “I FAILED the exam”; the word ‘failed’ may indicate strong negative or frustration emotion.
3. Remove punctuation: Removal of punctuation, such as question and exclamation marks, has been applied by many researchers, e.g. [92, 136, 190]. Punctuation does not usually hold any sentiment, however, there are situations when it can convey sentiment strength; for example, in “I passed!!!”, the exclamation mark may mean strong positive or joy emotion. In addition, question marks may represent confusion. Zhao et al. [190] and Wang et al. [181] did not remove punctuation due to them believing it held sentiment. On the other hand, most researchers removed all punctuation [92, 136].
4. Remove numbers: Removal of numbers is common in previous research, as usually they have no meaning; however, in chat applications, numbers can be used to represent words. For example, ‘to’ or ‘too’ can be written as ‘2’ and ‘for’ can be written as the number ‘4’. One example for numbers that may represent sentiment is the word ‘great’ that can be written in chat language as ‘gr8’. Despite the fact that some numbers may represent sentiment, in most cases they are eliminated.
5. Removing repeated letters: Spelling mistakes can be corrected by removing extra letters [64, 92, 123, 136, 181]. Some researchers kept two letters and removed the extra repeated letters [64, 166, 181]. Other researchers, such as Agarwal et al. [1], kept three letters and removed any extra ones.
6. Remove stop words: There is no agreed set of stop words that should be removed. Stop words can be words such as ‘a’, ‘the’ and ‘there’. Removal of stop words will help reduce the index space and improve response time [3]. Examples of researchers that removed stop words are Zhao et al. [190], Mouthami et al. [115], Claster et al. [32] and Prasad [136].

7. Stemming: Returning the word to the basic form; for instance the word ‘learning’ becomes ‘learn’. Examples of researchers that used stemming are Claster et al. [32], Gamon et al. [57] and Kechaou et al. [83].
8. Spelling check: Checking the misspelled words is also an important preprocessing technique due to the incorrect classification or overlooking misspelled words. Smart phones and social media websites nowadays offer spelling correction making spelling mistakes less common.
9. Negation: Negation changes the polarity of a word i.e. positive into negative or negative to positive. Negators can affect the sentence and make it the opposite meaning. Some examples of negators are ‘not’, ‘none’, ‘nobody’, ‘never’ and ‘nothing’ [159]. In Song et al. [156] a rule-based algorithm was created to identify the negation sentences. Pang and Lee [131] also detected negation in the sentence by adding NOT\_ to all the words between the negator and the following punctuation sign. Some examples of negators are shown in Table 4.1. There are some difficulties with detecting negation such as:
  - The location of the negation can affect the sentence; for instance including the word ‘not’ at the end of the sentence [111]; and
  - Negating the sentence when the negation appears before words; such as the word ‘wonder’: “No wonder everyone loves it” [111].

Table 4.1 Negators Examples

Negation	Example
Nobody	Nobody understands this lecturer.
None	None of the examples are understood.
Nothing	Just another lesson, nothing spectacular

In addition to the preprocessing techniques described above, there are different types of preprocessing techniques used according to the source of the data. For example, if the data was collected from Twitter or a social media network, it would need different preprocessing techniques from collecting it by paper, such as replacing hashtags and Twitter special characters such as emoticons. Some of the common Twitter-specific data preprocessing techniques are: removing hashtags [92], removing URLs [92, 126, 136, 190], removing retweets [190], identifying emoticons [1, 92, 136, 190], removing user mentions in tweets [126, 190], removing Twitter special characters [92] and slang/chat language handling [1].

All preprocessing techniques above have been used in most of previous studies in the educational domain [107, 123, 174]. Spelling check is a common technique used in the

educational domain and was employed by Ortigosa et al. [123] and Martin et al. [107]. This technique is usually used with data collected from social media that contains chat language.

Negation is another common preprocessing technique that affects polarity [182]. In the educational domain, it is found in studies such as Ortigosa et al. [123] and Martin et al. [107].

### 4.1.2 Features

Features give a more accurate analysis of the sentiments and detailed summarization of the results [186]. The most common features are n-grams [1, 63, 180] and POS (part-of-speech) tagging [131, 181]. Additionally, there are less common features related to the data source (i.e. Twitter-related or domain-based features). These are explained in the subsections below, which also includes an overview of related research using these features.

#### N-grams

N-grams are sequences of n-items from some text. Letters, syllables, or words are all n-gram items. The most frequently used n-gram items are words and the most popular n-grams are unigrams (one word), bigrams (two sequent words) and trigrams (three sequent words).

Research shows that there is no clear answer on which n-grams lead to the best performance [1, 63, 64, 180]. There are studies where unigrams lead to a better performance than bigrams and trigrams [1, 12, 27, 56, 123, 136, 180]. One study about Twitter sentiment analysis argued that bigrams decrease the performance [63]. In contrast, other researchers in the same field such as Pak and Paroubek [126] found that bigrams lead to a higher accuracy than unigrams.

#### Part-of-speech (POS) tagging

Part-of-speech tagging consists of categorising the word into lexical categories, which in the case of the English language are: 1) Noun, 2) Pronoun, 3) Adjective, 4) Verb, 5) Adverb, 6) Preposition, 7) Conjunction and 8) Interjection. Different categories and combinations of parts-of-speech have been used for sentiment analysis. Table 4.2 gives an overview of these categories and their corresponding research studies.

Khuc et al. [85] found that POS-tagging led to a good performance. In contrast, Go et al. [64], Pang and Lee [131] and Wang et al. [181] claimed that POS-tagging led to a poor performance.

In the educational domain, n-grams have been used by Troussas et al. [174], while other researchers have used POS-tagging, such as Ortigosa et al. [123] and Martin et al. [107]. Both n-grams and POS-tagging features led to a good performance in the educational domain,

Table 4.2 Part-Of-Speech studies

Part-of-speech used	Abbreviation	Research
Degree Adverbs	Adv	Tian et al. [54] and Song et al. [156]
Verb, Adjectives, Adverbs and Nouns	VAdjAdvN	Agarwal et al. [1]; Go et al. [63]; Khuc [85]; Ortigosa et al. [123] and Go et al. [64].
Verb, Adjectives and Adverb	VAdjAdv	Kumar and Sebastian [92]
Verb, Adjectives and Nouns	VAdjN	Zhao et al. [190] and Martin et al. [107]
Verbs and Adjectives	VAdj	Barbosa and Feng [12]

making it difficult to distinguish which features would be most suitable for our application. Consequently, there is a need to explore several features and their combinations for our purpose.

### Other features

Other features could be added to increase the models performance. Lexicon-derived features such as polarity averages or sums have been commonly used [102, 189]. Many researchers have used lexicon-based approach alone, which is the process of detecting words found in the lexicon, calculating the averages or sums of the polarities, then labelling the sentence according to the majority score (i.e. the highest polarity score). Adding lexicon-derived polarity averages to the machine learning approach has also been found to increase the models' performance [189].

There are many general lexicons such as SentiStrength and NRC word-emotion [1]. SentiStrength is a lexicon of 2310 sentiment words and word stems classified into positive and negative scores from 1 to 5 [165]. NRC word-emotion lexicon consists of 14000 words classified into positive and negative sentiment and eight emotions (i.e. anger, disgust, fear, happy, sadness, surprise, trust, and anticipation) [112]. In the educational domain, Munezero et al. [118] used lexicons to extract emotions from students' learning diaries by comparing each sentence in the students' diaries against the NRC word-emotion association lexicon [112, 113]. However, our main focus is on domain-based lexicon, which in our case is education. Utilising a domain-based lexicon has the potential to provide more accurate results due to the different word meanings across domains. There is no lexicon available for our purpose, therefore we created one, which is explained in more detail in section 4.3.2.

Other features that could be added to the model are Twitter-related features such as number of emoticons [67, 71] or number of hashtags [71]. Emoticons have been used in different ways in sentiment analysis. Some researchers used emoticons to determine the polarity of the sentence [1, 64], other have used them as features [107, 123]. Hashtags are another Twitter-related feature that can be a useful as they could be used to express sentiment, however, they could also be used to mark topics [1]. Hashtags have been used in previous research as a feature to give extra value or strength to the sentiment [67, 71, 140].

Other less common features which could be used in any domain are the number of punctuation signs, and specifically, question marks and exclamation marks which could hold some value to the sentiment. The number of repetitions of the question marks and exclamation marks does not hold much meaning, however, the repetition itself could indicate some sort of subjectivity [184].

Guha et al. [71] study about sentiment analysis on general Twitter used unigrams, lexicons, question and exclamation marks, emoticons and number of hashtags. They found that unigrams, question and exclamation marks, emoticons and number of hashtags led to an improved performance in their models, while lexicons led to a decreased performance.

Emotion labels and the time of the lecture have not been used before, but could be useful in a study related to students' feedback. Sarcasm is another feature that could be used in sentiment analysis. Chapter 6 presents the experiments conducted for detection of sarcasm from students' tweets. Table 4.3 shows a summary of previous studies with features that they used.

### 4.1.3 Machine Learning Techniques

The most popular machine learning techniques for sentiment analysis are Naive Bayes (NB) [17, 57, 63, 111, 129, 160], Multinomial Naive Bayes (MNB) [63, 170], Support Vector Machines (SVM) [17, 44, 63, 111] and Maximum Entropy (ME) [63, 111]. Complement Naive Bayes (CNB) [170], Sequential Minimal Optimization (SMO) [166] and Random Forest (RF) [170] are machine learning techniques that have been used in data mining research before, however, they have not been explored much in the field of sentiment analysis [166, 170]. Consequently, they may be promising in sentiment analysis for education. These techniques are reviewed in the following.

#### Naive Bayes (NB)

*Naive Bayes* uses a probabilistic model that originates from Bayes theorem. It assumes independence between events/features. Naive Bayes performs well in sentiment analysis,

Table 4.3 Summary of studies with features

Research	Domain	Feature(s)
Munezero et al. [118]	Education	Lexicon-based
Guha et al. [71]	Twitter	Unigrams Lexicon-based Question marks Exclamation marks Emoticons Number of hashtags
Pang and Lee [131]	Movie reviews	Unigrams
Barbosa et al. [12]	Twitter	Hashtags All Punctuation marks (i.e. full stops, commas, question and exclamation marks) Exclamation marks
Kouloumpis et al. [90]	Twitter	Unigrams Lexicon-based Emoticons

e.g., [131, 136], as it only needs a small amount of training data to estimate parameters. Additionally, it can deal with discrete and continuous attributes and is fast and incremental [14]. Datasets sometimes consist of uneven classes, meaning one class has more instances than the other. Naive Bayes does not work well with uneven class sets and Complement Naive Bayes addresses this problem [65, 138].

### **Multinomial Naive Bayes (MNB)**

*Multinomial Naive Bayes* is a specific case of a Naive Bayes classifier which uses a multinomial distribution for each of the features. Multinomial Naive Bayes estimates the conditional probability of a particular feature (word) given a class as the relative frequency of term  $t$  in documents belonging to class  $c$ . Similar to Naive Bayes, it is fast, easy to implement and relatively effective [138].

### **Complement Naive Bayes (CNB)**

*Complement Naive Bayes* estimates feature probabilities for each class based on the complement of that class (i.e. based on the instances from all the other classes rather than the target class). The Complement Naive Bayes classifier improves the weaknesses in the Naive Bayes classifier (i.e. dealing with uneven classes). Research by Gokulakrishnan et al. [65]

about general Twitter sentiment analysis shows that Complement Naive Bayes gives a higher performance than Naive Bayes when the classes are uneven.

### **Support Vector Machines (SVM)**

A Support Vector Machine (SVM) is a non-probabilistic classifier. Models that include just positive and negative classes are classified using a binary classifier, while models with the neutral class included are classified using a multi-class classifier. SVM finds hyperplanes that separate the classes [91]. SVM is highly effective at traditional text categorisation and generally outperforms Naive Bayes [14].

The effectiveness of the SVM partially depends on the kernel [28]. There are different types of kernels, of which the most popular are linear, polynomial and radial basis functions.

The linear kernel usually works best with larger amounts of data. The linear kernel is graphed as a straight line. In contrast, non-linear kernels are graphed as curved lines and paths, and generally give better results [28]. The polynomial and radial basis functions are two types of non-linear kernels. The polynomial kernel is popular with natural language processing [28, 66] and the radial basis kernel is popular for sentiment analysis applications [28]. For our experiments, the LibSVM model [28] in Weka [183] was used.

### **Maximum Entropy**

The Maximum Entropy (ME) classifier is similar to the Naive Bayes classifier, except that instead of the features acting independently, the model finds weights for the features that maximize the likelihood of the training data using search-based optimization [131].

One advantage of ME is that it does not make assumptions about the relationships between features [131]. Consequently, in contrast to NB and SVM, it could potentially perform better when conditional independence assumptions are not met [14]. One drawback is that it is not very realistic in many practical problems, as real datasets contain random errors or noise, which create a less clean dataset [44]. Consequently, we expect that higher levels of preprocessing would lead to better results in the ME classifier.

### **Sequential minimal optimization (SMO)**

*Sequential minimal optimization* is an algorithm derived from SVM. It addresses the quadratic programming problem (i.e. the problem of optimizing a quadratic function of several variables subject to linear constraints on these variables) that arises during the training of support vector machines.



SMO utilizes the smallest possible quadratic programming optimization problem, which are addressed quickly and analytically, improving its scaling and computation time [133]. SMO does not require extra matrix storage and it is less susceptible to numerical precision problems [133]. Additionally, SMO performs well for linear SVMs because its computation time is controlled by SVM evaluation which can be expressed as a single dot product, rather than a sum of linear kernels [133]. Moreover, SMO performs well for SVMs with sparse inputs and for non-linear SVMs [133]. This is due to the reduced kernel computation time speeding up SMO.

### **Random Forest (RF)**

*Random Forest* is based on Random Tree classifier, and uses a number of decision trees to improve the classification rate. Random forests leave one third of the sample training set out of the trees set. When each tree is built, the data is run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. Eventually, the proximities are normalized by dividing by the number of trees. Proximities are used in locating outliers, producing low-dimensional views of the data, and replacing missing data. The advantages of RF are that it does not over-fit due to its capability of handling a large amount of data, it is also fast and memory efficient [18].

In the education domain, Naive Bayes is found to be the best technique by Troussas et al. [174]. Contrarily, Song et al. [156] identified Support Vector Machines as the best performing classifier. Thus, machine learning techniques can perform differently within the same domain. This prompts the need to investigate different techniques.

#### **4.1.4 Neutral Class**

The neutral class is useful for real life applications due to the nature of sentiment sometimes being neutral and excluding it forces instances into other classes (i.e. positive or negative) [1, 63, 180]. Some researchers omitted the neutral class because of the insufficient amount of neutral training data which leads to poor performance [126, 180].

In the educational domain, the neutral class is needed [107, 123]. Students' feedback can be neutral and neglecting the neutral class does not show a complete picture of the class opinion. It may distort the true proportions of the positive and negative opinions, when the neutral proportion is not known. Most of the researchers have included the neutral class when analysing data from the educational domain due to its usefulness [107, 123]. On the other hand, others have not included it, such as Kechaou et al. [83]. Therefore, the role of

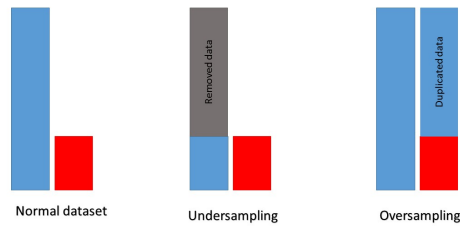


Fig. 4.1 Undersampling and Oversampling methods

neutral class needs to be investigated further for the purpose of our application. More about this is described in the results of section 4.2.

#### 4.1.5 Dealing with unbalanced datasets

There are many methods to deal with unbalanced classes in a dataset. The most common methods are undersampling and oversampling, which are illustrated in Figure 4.1. To clarify these two methods in a 2-class dataset, unbalanced classes consist of one class with more samples (i.e. the majority class) and another class with fewer samples (i.e. the minority class).

The undersampling method reduces the number of majority class members in the training set. In contrast, the oversampling method aims to increase the number of minority class members in the training set. More about this is explained in results of section 4.3.

#### 4.1.6 Evaluation metrics

All the models were tested using 10-fold cross-validation; accuracy, precision, recall and F-score were calculated. These were explained in chapter 3 in section 3.3.1.

The literature review outlined a diversity of preprocessing methods, features and machine learning techniques used for sentiment analysis. In the following experiments we explore all these aspects to identify their suitability for sentiment analysis of students' feedback.

## 4.2 Exploring N-grams and POS-Tagging features

In this study, we experiment with three of the most popular machine learning techniques, including Naive Bayes, Maximum Entropy and Support Vector Machine (SVM) with three

different kernels. In addition, we explore the Complement Naive Bayes classifier which is less common in sentiment analysis. Moreover, we also experiment with 12 different combinations of features and tested six preprocessing levels, including techniques such as negation, stemming and removal of stop words. In addition, we look at the effect of using the neutral class in all the models. For this experiment, we investigate the following aspects:

- what preprocessing techniques are the most effective and which level leads to the best results?
- which feature(s) are most suitable for this data?
- which machine learning techniques give the highest performance?
- what is the effect of including the neutral class in the models?
- considering the answers to the questions above, which is the optimal model for sentiment analysis of students' classroom feedback?

### 4.2.1 Data Corpus

Different sources were used for data collection. The first source was real-time feedback from lectures in the computing department at the University of Portsmouth. This included both postgraduate and undergraduate students. The students submitted feedback in textual form expressing their opinions and feelings about the lecture.

We also collected end of unit feedback from a mixture of institutes for various subjects. The total amount of data was 1036 instances of feedback (one from each student). By observing our dataset we found that it contains sentences written in formal language. Only 344 misspelled words were found from 24422 words in total, which is 1.41% of the data. The word errors were mainly missed spaces between words. The distribution of the feedback according to its source is presented in Table 4.4. The number of instances per class label is presented in Table 4.5. Examples of feedback from students for each of these classes are displayed in Table 4.6.

Table 4.4 Data Sources

Data Source	Number of instances
End of Unit Other Institutes	768
End of Unit University of Portsmouth	117
Real-time feedback University of Portsmouth	151

Table 4.5 Distribution of sentiment labels in our corpus

	Positive	Negative	Neutral
Frequency	641	292	103

The labelling of the data was done by three experts from linguistics and sentiment analysis backgrounds. The experts were asked to categorise each instance into one of the categories: positive, negative or neutral. The labels were chosen using the majority rule. When each expert chose a different label, the neutral label was assigned.

The reliability of the labels was verified using inter-rater reliability from which the percent agreement was 80.6%, the Fleiss kappa [55] was 0.625 and Krippendor's alpha [75] was 0.626. The percent agreement is considered over-optimistic, while the other two measures are known to be more conservative [101].

#### 4.2.2 Sentiment Analysis Model for Students' Feedback: N-gram and POS-tagging features

Our dataset is fairly clean because the feedback was written in standard English language, hence no chat language was used. Table 4.7 shows a description of all the preprocessing methods used. We experimented with six different level of preprocessing to ensure that the data was not over preprocessed, i.e. eliminating information that holds sentiment. These levels included different techniques that were chosen due to their popularity in previous studies:

1. Preprocessing P1: This is to test how well the model performs without any preprocessing apart from tokenization and converting the letters into lowercase. We included this to explore the role of different degrees of preprocessing, in which this one (P1) is the baseline;
2. Preprocessing P2: This is to remove numbers, punctuation, spaces and blanks and special characters. In most cases, these do not hold value or sentiment; therefore, they are noise to the data;
3. Preprocessing P3: This includes all of preprocessing P2 with an additional two techniques which are replacing 'n't' with not (e.g. replacing 'don't' with 'do not') and removing repeated letters. These two are fairly important steps to preprocess the data as it increases the probability of matching words to their right sentiment;

Table 4.6 Polarity Examples from Data Corpus

Polarity	Examples
Negative	<ul style="list-style-type: none"><li>• Maths is confusing.</li><li>• Sometimes communication with lecturers is difficult as some people have a language barrier that leads to confusion.</li><li>• Some units can be dull. Some marking not consistent.</li><li>• Some lecturers are not very good, they seem to just read off slides and are not helpful. Some of these lecturers when marking work seem inconsistent.</li></ul>
Positive	<ul style="list-style-type: none"><li>• Lectures well organised and clear readings easy to find.</li><li>• (Very!!) enthusiastic lecturer obviously knows the field very well is passionate about his subject which makes it interesting.</li><li>• I wouldn't change anything in the lecture. I like the way our lecturer engages everybody in the class, he makes sure everybody is listening and keeps everyone on their toes.</li><li>• Very informative and probably will be helpful in the coursework.</li></ul>
Neutral	<ul style="list-style-type: none"><li>• Somewhat difficult but interesting course.</li><li>• The lecturer was enthusiastic but at times seems intimidating to approach and ask questions.</li><li>• The sections about organ donation and copyright were really interesting but I found the externalities and cost-benefit lectures harder to follow.</li><li>• I understand the content. It was a lot to take in however, I knew a lot of this from other education.</li></ul>

Table 4.7 Preprocessing levels and techniques

Techniques	Preprocessing Level					
	P1	P2	P3	P4	P5	P6
Tokenization	Yes	Yes	Yes	Yes	Yes	Yes
Convert to lower case	Yes	Yes	Yes	Yes	Yes	Yes
Remove numbers	-	Yes	Yes	Yes	Yes	Yes
Remove Punctuation	-	Yes	Yes	Yes	Yes	Yes
Remove spaces and blanks	-	Yes	Yes	Yes	Yes	Yes
Remove special characters	-	Yes	Yes	Yes	Yes	Yes
Replace 'nt with not	-	-	Yes	Yes	Yes	Yes
Replace repeated letters with two	-	-	Yes	Yes	Yes	Yes
Remove stopwords	-	-	-	Yes	Yes	Yes
Stemming	-	-	-	-	Yes	Yes
Negation	-	-	-	-	-	Yes

4. Preprocessing P4: This includes all of preprocessing P3, plus the removal of stop word for instance 'the' and 'a'. This step removes all words that are irrelevant to the analysis [144];
5. Preprocessing P5: This includes all the preprocessing techniques in P4 plus stemming;
6. Preprocessing P6: This is the highest level of preprocessing. It includes all the previous preprocessing levels plus negation. The method we used to implement negation was the same approach as Pang and Lee [131], which consists of adding NOT\_ to each of the words between the negator and the following punctuation sign.

We chose these combinations according to the popularity and complexity of the preprocessing techniques. The inclusion in the different levels was done in order of complexity, starting from basic techniques such as tokenization, up to advanced techniques such as negation. To the best of our knowledge, our work is the first systematic experimentation with different levels of preprocessing aiming to identify what level is the most appropriate for our data, i.e. students' feedback, in combination with particular machine learning techniques.

In this experiment, we explore different n-gram combinations from unigrams, bigrams and trigrams. We also experiment with the most common POS-tagging features which are presented in Table 4.2.

The machine learning techniques used in our experiments are Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME) and Support Vector Machines (SVM) with different kernels (i.e. Radial Basis (RB), Polynomial (POL) and Linear (LIN)). CNB is rarely used in sentiment analysis; however, we use it to further test its potential in solving the uneven class problem for our application.

Table 4.8 Abbreviations and number of features

Term	# of Features
Unigrams (UNI)	3469
Bigrams (BI)	15944
Trigrams (TRI)	22338
Unigrams combined with bigrams (UNI+BI)	19413
Unigrams combined with trigrams (UNI+TRI)	25807
Bigrams combined with trigrams (BI+TRI)	38282
Unigrams combined with bigrams and trigrams (UNI+BI+TRI)	41751
Adverbs alone (Adv)	343
Verbs, Adjectives, verbs and Noun (VAdjAdvN)	3775
Verbs, Adjectives, Adverbs (VAdjAdv)	1594
Verbs, Adjectives and Nouns (VAdjN)	2881
Verbs and Adjectives (VAdj)	406
Neutral (Neu)	.....
Without Neutral (W/O)	.....

The number of features in our datasets for each feature combination are displayed in Table 4.8. The results obtained in these experiments are presented and discussed in the following section.

### 4.2.3 Results

The best machine learning results for each machine learning technique – with the corresponding preprocessing levels, features and presence of the neutral class – are displayed in Table 4.9.

Table 4.9 Machine learning techniques results

	NB	CNB	RB	POL	LIN	ME
Neutral or Without	W/O	W/O	W/O	W/O	W/O	W/O
Feature	Uni+Bi	Uni+Bi	Bi	Uni	Uni	Uni+Bi
Preprocessing	P4	P1	P4	P1	P1	P4
Accuracy	0.94	0.93	0.77	0.69	0.94	0.92
Precision	0.94	0.94	0.81	0.47	0.95	0.93
Recall	0.94	0.93	0.77	0.69	0.94	0.92
F-score	0.94	0.93	0.73	0.56	0.94	0.92

While these results are discussed in more detail in the following subsections, they also convey several high level observations as in the following:

- The best performance was obtained when higher levels of preprocessing were used, with the majority of the models performing best from level P3 to level P6 (the highest preprocessing level);
- Different classifiers performed best with different features; among the features that led to high classification performance are unigrams and unigrams combined with bigrams;
- From the machine learning techniques, SVM linear achieved the best performance with an accuracy of 94%. NB came in second place giving an accuracy of 94% and a slightly lower precision. SVM polynomial resulted in the lowest accuracy score at 69% with a precision of 47%;
- As can be seen from Table 4.9, the best performing models are without the neutral class; however, NB, CNB, LIN and ME still achieved a good performance when the neutral class is considered, with accuracies of 90%, 89%, 93% and 85%, respectively (see Table 4.20 further in this section 4.2.3).

### Preprocessing results

We looked at the performance difference between the models for all the levels of preprocessing for both n-grams and POS-tagging features to identify the best level for both features. For the experiments, we took an average of the accuracies for each of the n-gram models and the POS-tagging models for each level of preprocessing.

We observed from the results of the n-gram features, illustrated in Table 4.10 that the performance of most models increased at the highest level of preprocessing. However, CNB (with neutral) had its best performance at level P5, indicating that negation is not a good preprocessing technique for this machine learning technique. In contrast, most of the models showed a decrease in performance at level P5, suggesting that stemming does not perform well with n-grams. ME without the neutral class performed best without preprocessing. This is surprising as we originally anticipated that ME would perform best with a higher level of preprocessing. The differences between some of the levels are 0, indicating that in these cases, preprocessing makes no difference. Level P3 made no difference or decreased the performance in most of the models. This implies that with n-grams, removing repeated letters and changing 'n't' to 'not' is not useful.

As for the preprocessing difference in relation to the POS-tagging, illustrated in Table 4.11, the models performance varied. CNB-W/O, RB, POL-W/O and ME-W/O models showed their best performance with the highest preprocessing level. Other models such as NB and POL-Neu highest performances were at level P5. This suggests that negation



Table 4.10 Performance differences between preprocessing levels for n-grams (average across all n-gram models)

Level difference	NB-Neu	NB-W/O	CNB-Neu	CNB-W/O	RB-Neu	RB-W/O	POL-Neu	POL-W/O	LIN-Neu	LIN-W/O	ME-Neu	ME-W/O
P2-P1	-0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	-0.002	-0.001	0.000	-0.001
P3-P2	0.001	0.000	0.000	0.000	0.006	0.010	0.000	0.000	0.000	-0.002	0.000	-0.002
P4-P3	0.015	0.009	-0.032	-0.053	0.011	0.013	0.000	0.000	-0.020	-0.018	-0.012	-0.016
P5-P4	-0.127	-0.097	0.025	-0.040	-0.022	-0.048	0.003	-0.009	-0.088	-0.084	-0.023	-0.040
P6-P5	0.012	0.005	-0.179	0.051	0.018	0.034	0.000	0.010	-0.038	0.003	-0.053	-0.039

does not perform well with POS-tagging in these models. In CNB-Neu and ME-Neu, the highest preprocessing was obtained at level P3, indicating that negation, stemming and stop words do not perform well with POS-tagging in these models. The LIN-W/O models highest performance was at preprocessing level P2. Similar to the n-grams, in many models, the differences between some of the levels are 0. As outlined earlier, this could infer that the preprocessing techniques in those levels make no difference to the models. We notice that preprocessing level P4 decreased the performance of most of the models when the POS-tagging features were used, implying that removing stop words with POS-tagging removes valuable information.

Table 4.11 Performance differences between preprocessing levels for POS-tagging (average across all POS-tagging models)

	NB-Neu	NB-W/O	CNB-Neu	CNB-W/O	RB-Neu	RB-W/O	POL-Neu	POL-W/O	LIN-Neu	LIN-W/O	ME-Neu	ME-W/O
P2-P1	-0.001	0.007	0.019	0.016	-0.001	0.000	-0.001	0.000	0.000	0.011	0.005	0.013
P3-P2	0.002	0.001	0.001	0.011	0.000	0.000	0.000	0.000	-0.002	-0.001	0.002	0.000
P4-P3	-0.002	-0.003	-0.013	-0.041	0.000	0.000	0.000	0.000	-0.007	-0.008	-0.007	-0.009
P5-P4	0.011	0.002	-0.054	-0.012	0.003	0.004	0.001	0.000	-0.049	-0.038	-0.005	-0.013
P6-P5	-0.010	-0.013	-0.014	0.021	0.016	0.045	0.000	0.016	-0.015	-0.006	-0.010	0.011

To get a complete picture of the preprocessing levels, we calculated the performance difference for all the models between the preprocessing levels. The average of all the models accuracies across each level were taken; these are presented in Table 4.12. We found that the best preprocessing level is P6, except for NB-W/O, CNB-Neu, LIN and ME models. As for the NB-W/O and CNB-Neu models, the highest performance was obtained in level P4, indicating that stemming and negation decrease the performance of these models. Similar to the n-grams preprocessing difference results, the LIN-Neu showed its best performance with the lowest level of preprocessing. Moreover, similar to the POS-tagging preprocessing difference results, the LIN-W/O showed its best performance at P2. Likewise, the ME-W/O model showed its highest performance at level P2, while ME-Neu model showed its best performance at level P3.

Table 4.12 Performance differences between preprocessing levels for all features (average across all models)

	NB- Neu	NB- W/O	CNB- Neu	CNB- W/O	RB- Neu	RB- W/O	POL- Neu	POL- W/O	LIN- Neu	LIN- W/O	ME- Neu	ME- W/O
P2-P1	-0.001	0.004	0.010	0.008	0.000	0.000	0.000	0.000	-0.001	0.005	0.003	0.006
P3-P2	0.002	0.000	0.001	0.005	0.003	0.005	0.000	0.000	-0.001	-0.002	0.001	-0.001
P4-P3	0.007	0.003	-0.022	-0.047	0.005	0.000	0.000	0.000	-0.014	-0.013	-0.010	-0.013
P5-P4	-0.058	-0.048	-0.015	-0.026	-0.009	-0.022	0.002	-0.004	-0.069	-0.061	-0.014	-0.026
P6-P5	0.001	-0.004	-0.096	0.036	0.017	0.040	0.005	0.013	-0.026	-0.001	-0.031	-0.014

Some of the preprocessing levels improved the performance noticeably. Examples of these are presented in Table 4.13. For instance 'P4-P5' means that the performance of model using level P4 preprocessing is higher than the same model with level P5 preprocessing. One example is CNB-Neu with Adv, P6 increased the accuracy by 15% in comparison with P4, indicating that stemming and negation are good features with this model.

On the other hand, in some models, the lower levels of preprocessing led to a higher accuracy. For example, in the LIN-Neu model with trigrams, the highest accuracy was found in P4, and when applying level P5, the performance decreases by 46%. Likewise, in the ME-Neu model with unigrams, the best performance was with P1. When applying preprocessing P3, the models performance lowers by 27%. As mentioned earlier, this contradicts what we originally believed regarding the ME classifier performing best with a higher level of preprocessing.

We noticed more significant differences between preprocessing levels with the n-gram features than POS-tagging ones. This could be due to the reduced number of features that are selected with POS-tagging.

## Features results

The models with the n-gram features performed slightly better than POS-tagging. The following subsections outline the results of our experiments with n-grams and POS-tagging features.

### • *N-gram results*

The n-grams feature led to the best performance, which is similar to some previous research [1, 12, 27, 56, 123, 136, 180], but also in contradiction with other research [64], where POS-tagging models led to a better performance than n-grams models. The highest accuracy for each of combination of the n-gram features are presented in Table 4.14.

Table 4.13 Noticeable preprocessing level differences

Model	Feature	Preprocessing levels	Difference
NB-W/O	Uni	P4-P5	14%
	Uni+Bi	P4-P5	15%
	Uni+Bi+Tri	P4-P5	15%
NB-Neu	Uni	P4-P5	21%
	Uni+Bi	P4-P5	22%
	Uni+Bi+Tri	P4-P5	21%
CNB-Neu	Uni	P4-P5	14%
	Adv	P1-P5	19%
CNB-W/O	Adv	P1-P5	21%
	VAdjAdvN	P6-P4	15%
LIN-W/O	Uni	P4-P5	18%
	Uni+Bi	P4-P5	15%
	Uni+Tri	P4-P5	21%
LIN-Neu	Tri	P5-P4	46%
ME-Neu	Uni	P1-P3	27%
	Uni	P4-P5	19%
ME-W/O	Uni	P4-P5	15%
	Tri	P4-P5	18%
	Uni+Bi	P4-P5	15%
	Uni+Tri	P4-P5	13%
	Bi+Tri	P3-P5	24%
	Adv	P1-P5	16%

Examples of frequent unigrams commonly found in our dataset are illustrated in Table 4.15. Some of the interesting unigrams are: ‘technology’, ‘blackboard’, ‘examples’, ‘helpful’, ‘learning’, ‘questions’, ‘knowledge’, ‘topics’, ‘notes’, ‘lecturer’ and ‘subject’.

We found that bigrams and trigrams led to a good performance with the CNB model, implying that some 2-word and 3-word combinations have a high informative value for our particular application. One example that illustrates the usefulness of bigrams in our application is the bigram “more examples”. This can indicate that students are confused, and therefore the sentiment is negative. On the other hand, when the word ‘less’ is replaced with ‘more’, it may mean that the students are satisfied; thus the sentiment is positive. Bigrams usually have stronger sentiment when using words such as ‘too’, ‘very’ and ‘really’ before other words. Some examples for these are: “really interesting”, “very engaging”, “too early”. Table 4.16 shows the ten most frequent bigrams.

Table 4.14 N-gram results

	Uni	Bi	Tri	Uni+Bi	Uni+Tri	Bi+Tri	Uni+Bi+Tri
Neutral or without	W/O	W/O	W/O	W/O	W/O	W/O	W/O
MLT	LIN	RB	ME	LIN	NB	CNB	NB
Preprocessing	P1	P4	P1	P1	P4	P1	P4
Accuracy	0.94	0.77	0.71	0.94	0.93	0.90	0.93
Precision	0.95	0.81	0.76	0.95	0.94	0.91	0.94
Recall	0.94	0.77	0.71	0.94	0.93	0.9	0.93
F-score	0.94	0.73	0.62	0.94	0.93	0.9	0.93

Table 4.15 Frequent unigram words

Unigram	Frequency
lecture	182
use	151
class	140
maths	131
time	108
course	98
like	92
interesting	82
students'	73

Our research is similar to Wang and Wu [180] study about general Twitter sentiment analysis, where trigrams led to a poor performance. However, certain 3-word combinations have a high informative value for our particular application. Examples of three word combinations that commonly appeared in our dataset are shown in Table 4.17.

The best performance when using the unigrams combined with bigrams was obtained by the SVM linear model. This is similar to the research of Go et al. [63] about general Twitter sentiment analysis. Furthermore, Pang and Lee [131] research showed that unigrams combined with bigrams led to a good performance with the SVM classifier.

Unigrams combined with trigrams led to a good performance in the Naive Bayes model giving an accuracy of 93%.

Bigrams combined with trigrams are rarely used in sentiment analysis. This could be due to the complication of these two features combined. We found that this feature led to a good performance in the CNB model. Similarly, unigrams combined with bigrams and trigrams are rarely used as it is also complicated and takes a long time to run. Our

Table 4.16 Frequent bigrams

Bigram	Frequency
the lecture	78
the course	46
the lecturer	43
the subject	38
the board	34
have learnt	32
the student	32
very good	31
very interesting	23
to learn	19

Table 4.17 Frequent trigrams

Trigram	Frequency
learned a lot	18
on the board	17
of the lecture	17
maths can be	16
understand the content	9
learnt lots of	8
the best faculty	8
in the course	8
in the classroom	7
would be good	7

results show that the best performance when using this feature was obtained by the NB classifier.

- ***POS-tagging results***

POS-tagging features led to a good performance in several of the models; these are presented in Table 4.18. The Adverbs alone led to the highest accuracy at 88% with the CNB classifier. This is in contrast to the research by Tian et al. [54] in the e-learning domain, where using Adverbs led to the lowest performance.

The best performance when using the Verbs, Adjectives and Adverbs (VAdjAdv) features was obtained by the CNB classifier. Kumar and Sebastian [92] proposed sentiment analysis models including the VAdjAdv features, however, their study was only a case study and they did not validate their results or evaluate the performance.

Table 4.18 POS-tagging results

	Adv	VAdjAdvN	VAdjAdv	VAdjN	VAdj
Neutral or Without	W/O	W/O	W/O	W/O	W/O
Highest Technique	CNB	ME	CNB	LIN	LIN
Highest Preprocessing	P1	P6	P2	P4	P2
Accuracy	0.88	0.82	0.80	0.83	0.78
Precision	0.80	0.88	0.88	0.85	0.80
Recall	0.76	0.86	0.81	0.90	0.91
F-Score	0.78	0.87	0.85	0.87	0.85

Table 4.19 Best classifiers-features combinations

Classifier	SVM linear	CNB
Features	Uni	Bi+Tri
	Uni+Bi	Adv
	VAdjN	VAdjAdv
	VAdj	-

The ME classifier led to the best performance when the VAdjAdvN features were used. This is similar to Go et al. [64] research where they obtained a good performance when using the VAdjAdvN features with the ME classifier.

The best performance for all the other POS-tagging features was obtained by the SVM linear classifier.

### Machine learning results

Our results presented in Table 4.9 (at the beginning of this section) illustrate that all the classifiers performed well with a minimum of 77% accuracy, a maximum of 94% accuracy and an average of 86% accuracy from all the classifiers. The SVM linear and NB classifiers performed the best with accuracies of 94%.

CNB also gave a good accuracy at 93%. We found that SVM linear was the best classifier for 4 of the features, and CNB was the best for another 3 features – these are illustrated in Table 4.19. A more detailed discussion of the machine learning techniques results are presented below.

#### • *Naive Bayes results*

The Naive Bayes classifier performed well when using the n-grams with preprocessing levels P1 to P4. However, we found that negation and stemming decreased the results of NB noticeably in some instances by up to 22%. Moreover, Naive Bayes highest

performance was using unigrams combined with bigrams with the level P4 giving an accuracy of 94%.

Our results indicated that using the POS-tagging features with NB leads to a decreased performance compared with using the n-grams features. This is similar to many researchers where they found that NB performs poorly with POS-tagging [54, 123].

- ***Complement Naive Bayes results***

We found that Complement Naive Bayes addresses the problem of uneven training sets in Naive Bayes on our dataset. Our experiments show that Complement Naive Bayes performs much better than Naive Bayes in some of the models. Similarly, research by Gokulakrishnan [65] about analysing tweets had similar results. Our results indicate that CNB is a useful technique when there is only little training data for the neutral class which led to poor performance in previous research [63, 180]. We found that CNB performed best with the bigrams combined with trigrams, Adv and VAdjAdv features.

- ***Support Vector Machine results***

The SVM linear and radial basis models perform better with the n-grams than with the POS-tagging features. This is different from Song et al. [156] research about e-learning, where the SVM classifier with the POS-tagging features performed well with an 87% accuracy.

SVM polynomial had the highest accuracy using adverbs and with preprocessing level P6. It had the highest recall of all the models (97%), indicating a high sensitivity. On the other hand, SVM polynomial gave the lowest accuracy out of all the machine learning techniques, despite the fact that it is known to work well with natural language processing models. Similarly, research by Jain and Nayak [79] on movie reviews from IMDB found that the polynomial kernel performed the worse out of all classifiers; at the same time, they found the linear kernel to work best. We notice that using n-gram or POS-tagging makes nearly no difference to the SVM polynomial results.

In applications and domains such as customer feedback and movie reviews, SVM is known to perform fairly well [56, 83, 131]. In the educational domain, researchers such as Ortigosa [123] and Tian et al. [54] found that SVM performs poorly. On the other hand, research by Song et al. [156] about e-learning shows that SVM performs well. Our results show that SVM is the best classifier, indicating that when working with the educational domain, the context of learning and type of data has a significant bearing on the best performing models.

- **Maximum Entropy results**

For Maximum Entropy, the best performance found is with the unigrams combined with bigram features. ME was also the best classifier for the VAdjAdvN features. Our results indicate that ME performs better with n-gram features than with POS-tagging features. Our results are similar to the research by Pang and Lee [131], who found that ME with n-gram features performed better than with the POS-tagging features. We also noticed that ME gave a lower accuracy when using stemming and negation in the preprocessing.

### Neutral class results

Most of the models performed best without the neutral class, however, we notice that three classifiers, LIN, NB and CNB, still had high results when the neutral class was considered. The highest results for each classifier when the neutral class was included are presented in Table 4.20. SVM showed the best performance when using the neutral class with an accuracy loss of only 1% compared with the model without the neutral class.

Table 4.20 Highest performance for each classifier when the neutral class is included

	NB	CNB	LIN	POL	RB	ME
Feature	Uni+Bi	Uni+Bi	Bi	Uni	Uni+Bi	Uni+Bi
Preprocessing	P4	P4	P4	P1	P2	P4
Accuracy	0.90	0.89	0.72	0.62	0.93	0.85
Precision	0.92	0.90	0.66	0.38	0.93	0.84
Recall	0.90	0.89	0.72	0.62	0.93	0.85
F-score	0.91	0.89	0.65	0.47	0.93	0.82

By observing all the results, we found that the best models were obtained with the SVM classifier with different features. We found that the highest models were with unigrams, unigrams combined with bigrams and unigrams combined with bigrams and trigrams with preprocessing levels P1-P4 without the neutral class.

To distinguish how well these models performed against other classifiers, we used 6 paired t-tests (i.e. 36 experiments) with a significance level of 0.05 using Weka. We used t-test due to its capability of comparing different classifiers and finding if one classifier is more significant than the others. The annotation v or \* indicates that a specific result is statistically better (v) or worse (\*) than the baseline scheme, which in our case in the LIN classifier. As observed from the t-test results presented in Table 4.21, we found that LIN is statistically better than most of the classifiers in some of the models, however, in some cases



Table 4.21 T-tests for best models

	LIN	RB	POL	NB	CNB	ME
Features	Unigrams-W/O					
P1	93.78	68.7*	70.42*	87.99*	91.63	88.1*
P4	93.78	68.7*	75.78*	90.24	93.03	87.13*
Features	UNI+BI-W/O					
P1	94.42	68.7*	72.99*	90.45	91.53*	92.71
P4	94	68.7*	70.42*	92.28	93.56	91.1*
Features	UNI+BI+TRI-W/O					
P1	94.21	68.7*	71.06*	89.6*	91.53*	90.67
P4	92.92	68.7*	71.16*	92.07	93.35	91.21

it is equal to the accuracy of other classifiers such as in NB, CNB and ME, which give high accuracies.

As outlined above, the best overall performing model was the SVM linear model without the neutral class, using unigrams combined with bigrams and level P4 of preprocessing. The accuracy for this model is the highest out of all models at 94.42%, however, using other features with LIN also gives a 94% accuracy.

To reduce the complexity of the model and the time it takes for it to give an output, it is best to use as few features as possible; consequently, the optimal model would be using only unigrams. Despite the fact that this is the optimal model in terms of overall performance (accuracy, precision and recall), this model does not include the neutral class, which is essential for our application, as outlined in section 4.1.4. We found that the best performing model including the neutral class was also using the SVM linear classifier. The accuracy of this model is 93%, which is only 1% less than the model without the neutral class. This model also had a high precision and recall score: 93% and 93%, respectively.

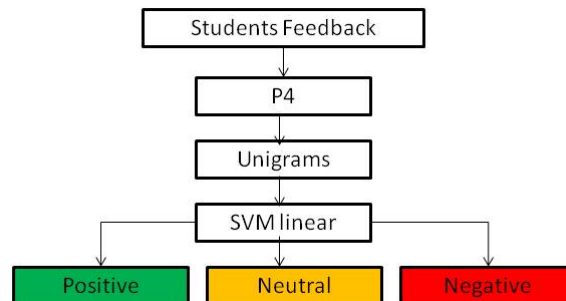


Fig. 4.2 Optimal combination of preprocessing, features and modelling technique for our application

To summarise, the best model found is using the SVM linear classifier with unigrams as features, using level P4 of preprocessing and including the neutral class. The Figure 4.2 shows the process of the model from the input to the output.

#### 4.2.4 Discussion

The results from the previous section indicated that preprocessing improves the performance of the different algorithms, with different levels working best for different algorithms, and that good results can be obtained with a variety of feature combinations. All algorithms performed better for the 2-class (positive/negative) problem compared with the 3-class (positive/negative/neutral) problem. This is not surprising, as the neutral class had the lowest number of instances, which makes it more difficult to predict. On the other hand, four of the six classifiers (NB, CNB, LIN and ME) performed well when the neutral class was included.

One limitation to this experiment is that the dataset could be considered to be relatively small, especially when comparing with domain-independent sentiment analysis studies. In domain-related studies, data may be more difficult to collect – in education, in particular, data collection in real settings is known to be difficult [77]. Another difficulty is the labelling of the data in a reliable way by human annotators, which is a time-consuming task; most research using large datasets have an automatic approach to labelling, mostly based on lexicons. Previous research, however, has been published where human annotated datasets were similar in size to or less than our dataset. Some examples of such studies covering both domain-dependent and domain-independent areas using human-labelled data are given in Table 4.22.

### 4.3 Exploring Other Features

In this experiment, we explore different features that are less common or have not been used before. Some of these features may be difficult to obtain in real-life systems such as the emotion feature, however they provide future researchers with the effect of different features on students' feedback polarity detection. The purpose of this study is to find the effect of these features on polarity detection to direct future researchers into exploring other features in other domains. We also experiment with less common machine learning techniques which are MNB, SMO and RF. Additionally, we investigate oversampling and undersampling to see their impact in dealing with unbalanced datasets.

Table 4.22 Examples of studies with relatively small datasets in both domain-independent (general) and domain-dependent areas.

Reference	Domain	Number of instances by class
Tan et al. [160]	Education reviews	Positive: 1,012; Negative: 254
	Stock reviews	Positive: 364; Negative: 683
	Computer Reviews	Positive: 544; Negative: 390
Pak and Paroubek [126]	General Tweets	Positive: 108; Negative: 75; Neutral: 33
Go et al. [63]	General Tweets	Positive: 182; Negative: 177
Thet et al. [167]	Film Reviews	Positive: 185; Negative: 185
Paltoglou et al. [128]	Cyberspace	Dataset1 – Positive: 41, Negative: 457, Neutral: 96 Dataset 2 – Positive: 107, Negative: 221, Neutral: 144

### 4.3.1 Data Corpus

The data collected was students' feedback expressing opinions and feelings about the lecture. They were from lectures taught in English in Jordanian universities on different topics: Calculus, English communication skills, Databases, Engineering, Molecular biology, Chemistry, Physics, Science, Contemporary history of the world and Architecture.

Twitter was used to collect the data. Students were asked to provide a polarity label (positive/ negative/ neutral) and choose one emotion from a set of 19 learning-related emotions provided, e.g. amused, bored, confused, frustrated for each tweet. More details about these 19 emotions and why they were chosen are presented in chapter 5. The total number of tweets collected were 1522 tweets with their corresponding emotion and polarity labels. The distribution of the polarity labels is found in Table 4.23. We found by observing the tweets, although the students were using Twitter they wrote in a formal manner.

Table 4.23 Distribution of sentiment labels in our corpus

	Positive	Negative	Neutral
Frequency	526	754	238

### 4.3.2 Sentiment Analysis Models for Students' Feedback: Other features

Three preprocessing levels using different techniques were tested to find the optimal level of preprocessing for the data. This includes a low level preprocessing, a medium level consisting of common general preprocessing techniques and a high level preprocessing consisting of Twitter preprocessing techniques as well. These were also chosen due to their popularity in previous studies [92, 126, 136, 190] and based on the findings from our first experiment in section 4.2:

1. Preprocessing P1: This is to test how well the model works without any preprocessing apart from converting the letters into lowercase. We included this to explore the role of different degrees of preprocessing, in which this one (P1) is the baseline;
2. Preprocessing P2: This is to remove numbers, punctuation, spaces and blanks and special characters. In most cases these do not hold value or sentiment, therefore they are noise to the data;
3. Preprocessing P3: In addition to the preprocessing in level P2, this includes the removal of hashtags, emoticons and Twitter special characters.

Findings from the previous study were used in the setup of this study. We used unigrams as features due to the results on the first experiment indicating that unigrams lead to the highest results in several classifiers. We also explored other features in combination with unigrams:

1. emotion label (i.e. amused, anxiety, appreciation, awkward, bored, confusion, disappointed, embarrassed, engagement, enthusiasm, excitement, frustration, happy, motivated, proud, relief, satisfaction, shame and uninterested.)
2. number of all punctuation characters (e.g. commas, semicolons, brackets, apostrophes, etc.)
3. number of question and exclamation marks
4. number of emoticons
5. number of hashtags
6. time of the tweet during the lecture, i.e. beginning, middle, end

7. sarcasm, i.e. sarcastic and non-sarcastic labels (more details about how we obtained these labels are found in chapter 6)
8. lexicon-based features, i.e. Lexicon-Positive, Lexicon-Negative, Lexicon-Neutral.

All combinations of these additional features with unigrams were explored. We converted the nominal features to numeric (i.e. either 1 or 0) in Weka. For example, the emotion labels which were originally nominal were converted into several numeric features, where each feature is equal to 1 if the sentence included the emotion and 0 if it did not. This setup allows the investigation of the usefulness of each emotion feature.

A lexicon is a dictionary of words with their percentages of polarity. An example of typical information from the lexicon we built is shown in Table 4.24. We created a domain-based lexicon using 3310 words found in previous students' feedback, which was labelled by three experts to positive, negative and neutral. The feedback was collected from a mixture of institutes for various subjects, this was described previously in section 4.2.1. The Positive/Negative/Neutral lexicons were created by counting the number of that word's appearance in each class (i.e. positive, negative and neutral) and then dividing it over total number of that word appearances.

Table 4.24 Lexicon Example

Word	Positive-lexicon	Negative-lexicon	Neutral lexicon
sir	0.500	0.000	0.500
class	0.366	0.208	0.426
maths	0.497	0.000	0.503
good	0.443	0.098	0.459
course	0.256	0.391	0.353
students	0.351	0.243	0.405
lecture	0.303	0.311	0.386
lecturer	0.273	0.424	0.303
teacher	0.368	0.189	0.443
session	0.485	0.015	0.500
subject	0.378	0.220	0.402
technology	0.133	0.300	0.567
blackboard	0.316	0.263	0.421

To illustrate how the lexicon information was built, the word 'class' is used as an example. We find that it appears 79 times in the positive feedback, 45 times in the negative, and 92 times in the neutral. The lexicon for the word 'class' becomes 0.3657 for the positive lexicon, 0.2083 for the negative, and 0.42592 for the neutral. These are 79, 45 and 92 divided over 216 (i.e. the total occurrences of the word 'class' across the 3 polarity labels).

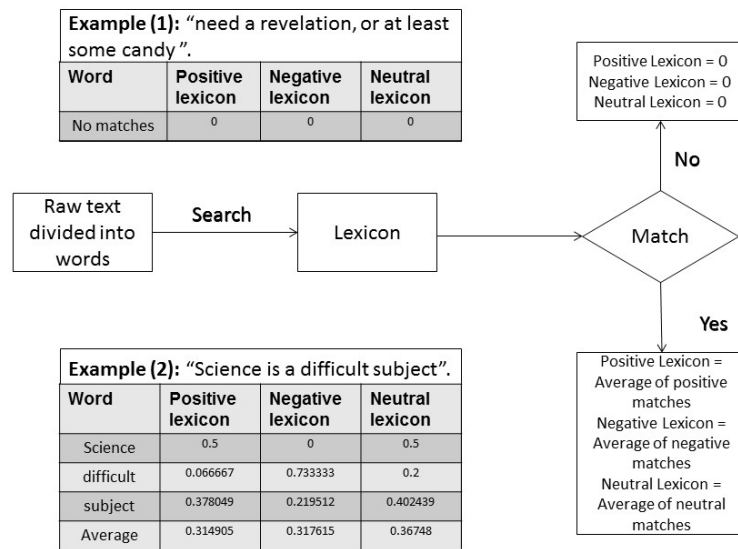


Fig. 4.3 Lexicon match example

The method that we used to calculate the lexicon-based features was to take the average of the Positive-lexicon, Negative-lexicon and Neutral-lexicon for the words found in a sentence that were detected in the lexicon. An illustration is shown in Figure 4.3. For instance, if we had a sentence "This was an interesting day", the lexicon count is illustrated in Table 4.25.

Table 4.25 Lexicon count example

	Positive-Lexicon	Negative-Lexicon	Neutral-Lexicon
Interesting	0.5	0	0.5
day	0.4574	0.0638	0.4787
Lexicon-based features	0.4787	0.03191	0.4893

As for the sarcasm feature, the data was labelled into two classes, sarcastic and non-sarcastic. This was done by looking for sentences that expressed one emotion but was labelled another emotion, and by looking for the word 'sarcasm' or 'sarcastic'. More information about how the data was labelled as sarcastic or non-sarcastic is presented in chapter 6.

The machine learning techniques we used were: Naive Bayes (NB), Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Maximum Entropy (ME), Support Vector Machines (SVM), Sequential Minimal Optimization (SMO) and Random Forest (RF), due to their popularity and promising results in previous research.

### Feature Usefulness

To find the feature usefulness, we applied 5 tests: Chi Square, Information Gain, Gain Ratio, Correlation and OneR. These tests were chosen due to their popularity in previous research [19, 81, 97]. A brief description of the tests is presented in the following:

- *Chi square* method evaluates features individually by measuring their chi-squared statistic with respect to the classes.
- *Information Gain* method evaluates the worth of a feature by measuring the information gain with respect to the class.
- *Gain Ratio* evaluates features individually by measuring gain ratio with respect to the class.
- *Correlation* evaluates features by measuring the correlation (Pearson's) between it and the class. The nominal attributes or features are considered on a value by value basis and an overall correlation for a nominal attribute is arrived at via a weighted average.
- *OneR* uses the minimum-error attribute for prediction, discretizing numeric attributes.

The results of the feature usefulness tests are presented in Appendix B. We found that across all the tests, there is a similarity in feature usefulness. Most of the features were found at the same level in feature usefulness or just one level place difference. Unlike the other tests, the OneR test resulted in a different ranking of the features according to usefulness. This could be due to its functionality in using the minimum-error attribute for prediction. As the results were relatively similar across tests (with the exception mentioned above), we present here the results from the Chi Square test. The ranking of the features are presented in Table 4.26 in the first two columns. We experimented with all the features in combination with unigrams (Uni+32), the top 15 features (Uni+15), the top 7 features (Uni+7) and the top 2 features (Uni+2). These were established by approximatively halving the number of features each time, with the purpose of finding the lowest number of features that still leads to a good performance. We chose the top 2 features (Bored and Amused emotions) due to them being the top two features across nearly all tests. We also experimented with lexicons alone, due to its popularity in previous research [102, 189].

Table 4.26 Features in each dataset

With Neutral	Without Neutral	UNI+2	UNI+7	UNI+15	UNI+32
Bored	Bored	Y	Y	Y	Y
Amused	Amused	Y	Y	Y	Y
Positive-lexicon	Positive-lexicon	N	Y	Y	Y
Negative-lexicon	Negative-lexicon	N	Y	Y	Y
Excitement	Neutral-lexicon	N	Y	Y	Y
Neutral-lexicon	Excitement	N	Y	Y	Y
Happy	Happy	N	Y	Y	Y
# of hashtags	Frustration	N	N	Y	Y
Frustration	# of hashtags	N	N	Y	Y
Sarcasm	Sarcasm	N	N	Y	Y
Engagement	Engagement	N	N	Y	Y
Enthusiasm	Enthusiasm	N	N	Y	Y
Embarrassed	Confusion	N	N	Y	Y
Confusion	Anxiety	N	N	Y	Y
Anxiety	Embarrassed	N	N	Y	Y
Shame	Appreciation	N	N	N	Y
Awkward	Awkward	N	N	N	Y
Disappointed	Motivated	N	N	N	Y
Motivated	Disappointed	N	N	N	Y
Relief	Proud	N	N	N	Y
Proud	Middle-lecture	N	N	N	Y
Satisfaction	End-lecture	N	N	N	Y
Appreciation	# of question and exclamation marks	N	N	N	Y
# of question and exclamation marks	# of emoticons	N	N	N	Y
# of emoticons	# of punctuation	N	N	N	Y
Middle-lecture	Relief	N	N	N	Y
End-lecture	Satisfaction	N	N	N	Y
# of punctuation	Shame	N	N	N	Y
Uninterested	Uninterested	N	N	N	Y
Beginning-lecture	Beginning-lecture	N	N	N	Y



### 4.3.3 Results

We included models with the neutral class and without. We had 144 models with the neutral and 144 models without the neutral class (Preprocessing levels (3) x Feature combinations (6) x Machine learning techniques(8)). We evaluated the models using 10-fold cross-validation and calculated the accuracy, precision, recall, F-score and error rate. The error rate is a metrics that is sometimes included in sentiment analysis which represents the percentage of instances that are not accurate. In addition, as our data was unbalanced, we experimented with undersampling and oversampling methods.

The best models for each machine learning techniques – with the corresponding preprocessing levels and features with the neutral class are displayed in Table 4.27 and without the neutral class in Table 4.28.

The baseline feature was unigrams which we compare other models to; we displayed the unigram results with the same preprocessing level as the best results found in Table 4.27 and Table 4.28, in Table 4.29 and Table 4.30 respectively. These results are discussed in more detail in the following.

Table 4.27 Best models with the neutral class

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI+15	UNI+15	UNI+32	UNI+15	UNI+32	UNI+15	UNI+32	UNI+15
Preprocessing	P1	P2	P1	P1	P2	P2	P2	P1
Accuracy	0.65	0.65	0.67	0.63	0.63	0.66	0.67	0.65
Precision	0.66	0.67	0.67	0.60	0.64	0.67	0.65	0.61
Recall	0.65	0.65	0.67	0.63	0.63	0.65	0.67	0.65
F-score	0.65	0.65	0.67	0.56	0.63	0.66	0.66	0.62
Error rate	0.35	0.35	0.33	0.37	0.37	0.35	0.33	0.35

Table 4.28 Best models without the neutral class

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15
Preprocessing	P1	P2	P1	P2	P3	P1	P3	P1
Accuracy	0.81	0.80	0.80	0.76	0.78	0.83	0.81	0.79
Precision	0.81	0.80	0.80	0.81	0.78	0.83	0.81	0.79
Recall	0.81	0.80	0.80	0.76	0.78	0.83	0.81	0.79
F-score	0.81	0.80	0.80	0.73	0.78	0.83	0.81	0.79
Error rate	0.19	0.20	0.20	0.24	0.22	0.17	0.19	0.21

Table 4.29 Unigram models with the neutral class

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI	UNI	UNI	UNI	UNI	UNI	UNI	UNI
Preprocessing	P1	P2	P1	P1	P2	P2	P2	P1
Accuracy	0.57	0.62	0.63	0.57	0.59	0.64	0.62	0.58
Precision	0.57	0.64	0.64	0.58	0.59	0.62	0.60	0.57
Recall	0.57	0.62	0.63	0.57	0.59	0.64	0.62	0.58
F-score	0.56	0.63	0.63	0.47	0.59	0.61	0.60	0.57
Error rate	0.43	0.38	0.37	0.43	0.41	0.36	0.38	0.42

Table 4.30 Unigram models without the neutral class

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI	UNI	UNI	UNI	UNI	UNI	UNI	UNI
Preprocessing	P1	P2	P1	P2	P3	P1	P3	P1
Accuracy	0.70	0.77	0.76	0.68	0.71	0.76	0.72	0.70
Precision	0.70	0.77	0.77	0.76	0.70	0.76	0.73	0.70
Recall	0.70	0.77	0.76	0.68	0.71	0.76	0.72	0.70
F-score	0.70	0.77	0.76	0.61	0.70	0.76	0.72	0.70
Error rate	0.30	0.23	0.24	0.32	0.29	0.24	0.28	0.30

Table 4.31 Performance differences between preprocessing levels models with neutral

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
P3-P2	-0.005	-0.028	-0.046	0.005	-0.008	-0.021	0.000	-0.010
P2-P1	0.002	-0.002	-0.001	-0.011	0.005	-0.002	-0.003	-0.001

### Preprocessing results

Table 4.27 represents the highest performance in models with the neutral class. We found the best preprocessing level is P2 which led to the highest performance in 4 models: MNB, LIN, ME and SMO. On the other hand, P1 led to the best performance in NB, CNB and RF models. According to Table 4.28 for the highest performance in models without the neutral, the best level of preprocessing was P1 across 4 models: NB, CNB, ME and RF. As for the MNB and RB classifiers, the highest preprocessing level was P2. LIN and SMO achieved their highest performances using preprocessing level P3.

To explore the performance difference between the models for all the levels of preprocessing, we took an average of the accuracies for each of the models (i.e. neutral and without) for each level.

We observed from the results of the models with the neutral, illustrated in Table 4.31 that the performance of most models was best at the lowest level of preprocessing. NB and LIN models had their best performance at level P2, indicating that removing numbers and punctuation is a good method for these models and increases their performances. We noticed that the RB model had its best performance at the highest level of preprocessing, however level P2 led to a decreased performance for this model. Most of the models showed decreased performances at level P3, indicating that hashtags, and emoticons might hold some value to the sentiment.

We observed from the results of the models without the neutral, as illustrated in Table 4.32, that the performance of models varied in the level of preprocessing. MNB, RB and LIN performed their best at level P2 of preprocessing. The difference between preprocessing level P2 and P1 in CNB and LIN are 0, indicating that in this case, preprocessing makes no difference. NB, ME and RF showed their best performance at preprocessing level P1, indicating that any preprocessing affects negatively on the performance, as mentioned earlier, hashtags, emoticons, numbers and punctuation may hold some value to the sentiment. Most of the models performance decreased when adding the highest level of preprocessing.

Table 4.32 Performance differences between preprocessing levels for models without neutral

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
P3-P2	-0.012	-0.030	-0.041	-0.013	0.000	-0.021	-0.003	-0.011
P2-P1	-0.005	0.001	0.000	0.004	0.003	-0.006	0.000	-0.008

Table 4.33 Difference between best models with the neutral class and unigram models

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI+15	UNI+15	UNI+32	UNI+15	UNI+32	UNI+15	UNI+32	UNI+15
Preprocessing	P1	P2	P1	P3	P2	P2	P2	P1
Accuracy	8%	3%	4%	7%	4%	2%	5%	7%
Precision	9%	3%	3%	4%	5%	5%	5%	4%
Recall	8%	3%	4%	7%	4%	1%	5%	7%
F-score	9%	2%	4%	10%	4%	5%	6%	5%
Error rate	-8%	-2%	-4%	-7%	-4%	-1%	-5%	-7%

## Feature results

In comparison with the performance of the unigrams feature alone, we found using UNI+15 led to the highest results. In Table 4.33, we calculated the differences between the highest results presented in Table 4.27 and the models with the unigrams feature- at the same level of preprocessing.

We can observe from the results that the Uni+15 dataset led to an increase in the results of unigrams by up to 8% in the NB model, and 7% in the RB and RF models. We found that CNB, LIN and SMO classifiers performed best using all the features (i.e. UNI+32). They led to a better performance than unigrams alone increasing the accuracy by 5% in the SMO model and 4% in both the CNB and LIN model.

The Uni+15 dataset included some of the emotion features which were: Bored, Amused, Excited, Happy, Frustrated, Engaged, Enthused, Embarrassed, Confused and Anxiety. This could be due to these were the most frequent emotion labels in the data except from Happy, which is surprisingly the third most useful feature according to the feature usefulness experiment in section 4.3.2. The frequencies of these emotions in our dataset can be found in chapter 5, section 5.2.1, Table 5.2.

Lexicon-based features do not make any difference in the performance of the models in comparison with unigrams alone. In contrast, researchers such as Lu et al. [102] and Zhang

Table 4.34 Difference between best models without the neutral class and unigram models

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15	UNI+15
Preprocessing	P1	P2	P1	P2	P3	P1	P3	P1
Accuracy	<b>11%</b>	3%	4%	8%	7%	7%	8%	9%
Precision	11%	3%	3%	5%	8%	7%	8%	9%
Recall	11%	3%	4%	8%	7%	7%	9%	9%
F-score	11%	3%	4%	12%	8%	7%	9%	9%
Error rate	-11%	-3%	-4%	-8%	-7%	-7%	-8%	-9%

et al. [189] found that using lexicon-based features led to a higher performance than when excluding them.

Lu et al. [102] research showed that unigrams combined with lexicons gave a better performance than unigrams alone, however, their study was based on Chinese data, therefore, it is difficult to compare with our results. Likewise, Zhang et al. [189] used unigrams combined with lexicons on Twitter data which led to a high performance.

The number of hashtags was found a useful feature according to the feature usefulness experiment in section 4.3.2. Guha et al. [71] on the other hand found hashtags not useful when using them with a lexicon-based sentiment analysis as it led to 57% accuracy which is relatively low.

Sarcasm was also a useful feature found in the UNI+15 features model. Sarcasm is an ongoing problem in sentiment analysis, and identifying the sarcastic sentences from the non-sarcastic can contribute to the polarity classification by making the detection more accurate.

We found that the least useful features according the experiment in section 4.3.2 were the least occurring emotion labels such as relief and proud, number of question and exclamation marks, number of emoticons, the time of the lecture and number of punctuation signs. In contrast to our study, Guha et al. [71] found the question and exclamation marks useful in their research on Twitter sentiment analysis.

As for the models without the neutral, using UNI+15 led to the highest performance in all the models. We can observe from Table 4.34 the difference in performance from the unigram models.

In general, the Uni+7 dataset led to a high performance than using unigrams alone and in some models the increase was 6%, however, using 15 features and all the features combined leads to a higher performance increasing some of the models performance up to 11%.

Table 4.35 Best classifier for each of the datasets with the neutral class

Feature	Uni	Uni+32	Uni+15	Uni+7	Uni+2	Uni+Lexicons
Best classifier	ME	CNB	SMO	SMO	CNB	CNB
Preprocessing	P1	P1	P1	P1	P1	P1
Accuracy	0.65	0.67	0.67	0.66	0.65	0.63
Precision	0.63	0.67	0.65	0.63	0.66	0.64
Recall	0.65	0.67	0.67	0.66	0.65	0.63
F-score	0.61	0.67	0.65	0.64	0.65	0.63
Error rate	0.35	0.33	0.33	0.34	0.35	0.37

Similar to the Uni+7 dataset, the Uni+2 dataset (i.e. emotion labels of bored and amused) led to a higher performance than the unigrams alone in some models up to a 5% increase.

### Machine learning technique results

The best machine learning technique for the models with the neutral class are presented in Table 4.35. We found that the best classifier was CNB which resulted in the highest performance for three of the models (i.e. Uni+32, Uni+2, Uni+Lexicons). This could be due to the unbalanced classes in the model with the neutral class – according to the results of the first experiment in section 4.2, CNB addresses the issue of unbalanced classes.

Our findings showed that SMO led to the highest performance in the Uni+15 and Uni+7 datasets. Likewise, Thelwall et al. [166] research showed that SMO led to a highest performance compared to the other classifiers, giving an accuracy of 58%, which is still relatively low.

As for the unigrams alone, we found that the best classifier was ME. This is similar to the research about movie reviews done by Pang and Lee [131], where the use of unigrams resulted in approximately 80% accuracy.

We noticed that the SVM classifier with both the RB and LIN performed the lowest. Similarly, Guha et al. [71] used SVM for sentiment analysis on Twitter, with features similar to ours (i.e. number of hashtags, presence of question and exclamation marks) and found relatively low results. Additionally, in the educational domain Ortigosa [123] also found that SVM gave the lowest results. In contrast many studies used SVM in combination of lexicons, and found that it performed well [102, 189].

As for the models without the neutral, the best classifiers according to each of the feature combinations are displayed in Table 4.36. As we can notice from the results, the best classifier is ME, leading to the highest in 4 feature combinations. The second best classifier was MNB, similarly to Go at al. [57] who found that MNB performed well when analysing tweets

Table 4.36 Best classifier for each of the datasets without the neutral class

Feature	Uni	Uni+32	Uni+15	Uni+7	Uni+2	Uni+Lexicons
Best classifier	MNB	ME	ME	ME	ME	MNB
Preprocessing	P1	P1	P1	P1	P1	P1
Accuracy	0.77	0.83	0.83	0.80	0.80	0.77
Precision	0.77	0.83	0.83	0.80	0.80	0.77
Recall	0.77	0.83	0.83	0.80	0.80	0.77
F-score	0.77	0.83	0.83	0.80	0.80	0.77
Error rate	0.23	0.17	0.17	0.20	0.20	0.23

Table 4.37 Oversampling the best models with the neutral class

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	15	15	15	15	15	15	15	15
Preprocessing	P1	P2	P1	P2	P3	P1	P3	P1
Accuracy	-7%	1%	0%	-9%	4%	-3%	2%	5%
Precision	-7%	0%	0%	-5%	4%	1%	4%	9%
Recall	-7%	1%	0%	-9%	4%	-2%	2%	5%
F-score	-8%	1%	0%	-11%	4%	-3%	3%	8%
Error rate	7%	-2%	0%	9%	-4%	3%	-2%	-5%

leading to an accuracy of around 81%. However, in contrast to our study, they found that ME leads to the lowest performance in comparison with other machine learning techniques.

### Dealing with unbalanced classes

Due to having an unbalanced dataset we experimented with undersampling and oversampling. We found that undersampling did not make any difference to the performance of the models and in some instances decreased the performance i.e. with the NB classifier. This could be due to undersampling removing information which could be valuable. On the other hand, oversampling keeps all the information in the training dataset.

To illustrate the performance of oversampling, we take the best models which were illustrated in Table 4.27 and oversample the data. The result of oversampling are presented in Table 4.37.

Oversampling had a different effect on different classifiers. The MNB, LIN, SMO and RF classifiers' performance improved with oversampling. The highest increase was found in the RF classifier, increasing models i.e. 5%. Similarly, Gokulakrishnan et al. [65] research on sentiment analysis of general Twitter claimed that using oversampling increases the performance of the classifiers (i.e. SMO, SVM and RF).

Oversampling led to a decrease in performance in the NB, RB and ME classifiers. The most noticeable decrease was in the RB classifier, decreasing the performance by 9%. Mountassir et al. [114] found that NB is not sensitive to unbalanced datasets, which probably could explain the decreased performance when oversampling. In the CNB classifier, oversampling made no difference to performance.

#### 4.3.4 Discussion

The following points summarise the findings of the experiments with other features:

- For the models with the neutral class, the best performance was obtained at preprocessing level P2, and in the models without the neutral class they were obtained with level P1;
- Unigrams with 15 features led to the best results for most of the models;
- From the machine learning techniques, CNB led to the best results for the datasets with the neutral, and ME led to the best results with datasets without the neutral;
- Undersampling the data decreased or did not have any effect on the performance of the models; and
- Oversampling improved some of the models.

In this experiment we investigated different features in the education domain and in particular for students' classroom feedback context that could improve the performance of polarity detection compared with using unigrams alone.

Although our results indicated that using UNI+15 led to the best performance, in reality this is very difficult to implement due to not having labels from participants for features such as emotion-based features. One way of capturing emotion labels is to ask the user to add an emotion label hashtag, however, if this is not supplied, then other methods could be investigated to detect emotion. We found that the two highest ranking features were the amused and bored emotion labels according to the usefulness experiment in 4.3.2. This could be due to high occurrence of these two emotions in the data. Detecting these emotions can contribute to an increase in accuracy (i.e. 5% in the dataset without the neutral).

The only features that could be calculated in practice from UNI+15 are Positive/ Negative/ Neutral lexicons and the number of hashtags. When experimenting with Positive/ Negative/ Neutral lexicons with unigrams, there was no significant increase or improvement to the models performance. Hashtags were explored only in combination with other features in the



UNI+15 dataset, therefore its influence in isolation (i.e. just unigrams with hashtags) is not known.

Other features that could be calculated are the number of question and exclamation marks, number of emoticons and number of punctuation marks. However, these features were found less useful according to the feature usefulness experiment in section 4.3.2.

We found that domain based features which are lexicons and time of lecture have a different effect in our application. We explored lexicons alone and found that it does not have any effect on performance, despite that lexicons were in the top 7 features. However, in combination with other features in the UNI+15 dataset, lexicons increase the performance. This could be due to the effect of other features in the UNI+15 dataset. As for the time of the lecture, we found it was not found useful according to the feature usefulness experiment in section 4.3.2.

Twitter-related features varied in usefulness, some more useful than others. The number of hashtags were found useful (in combination with other features as outlined above), which is surprising considering they could be used for other reasons not related to subjectivity. On the other hand, we found that using the number of emoticons is not useful, contradicting to many other studies which have found them useful in polarity detection.

Despite the findings in other studies that have shown the repetition of some punctuation marks could infer subjectivity [184], this experiment found that using punctuation features are not useful for our application.

Similarly to the first experiment in section 4.2, a limitation of this experiment is that the dataset is relatively small, in comparison with other sentiment analysis studies. However, generally data from domain-related studies are be more difficult to collect, particularly in the education domain and data from real settings [77]. In Table 4.22 in section 4.2.4, we found that in comparison with other sentiment analysis studies, our dataset is at an average size.

## 4.4 Optimal model for Polarity detection

Experiments 1 and 2 suggest that preprocessing at a lower level leads to a better performance than a higher level, as it may remove instances that are valuable to the sentiment.

The results of experiment 1 indicates that using unigrams leads to the highest performance. In contrast, experiment 2 suggests that using additional features in combination to the unigrams increases the performance. However in practice, some of these features are difficult to obtain. Therefore, unigrams may be the best feature in our application.

By observing the results in experiments 1 and 2, we found that SVM led to the best performing models in the experiment 1, however, in experiment 2, it led to the worst. We

found that CNB is the best choice for this application, due to its capability of dealing with unbalanced classes and its high performance with the neutral class.

In conclusion to the results of the experiments 1 and 2, we found that the best combination of preprocessing, features and machine learning techniques include the following:

- Preprocessing: a lower level preprocessing which includes converting text to small case and removing numbers and punctuation
- Features: unigrams
- Machine learning technique: CNB

An illustration of the models process is presented in Figure 4.4.

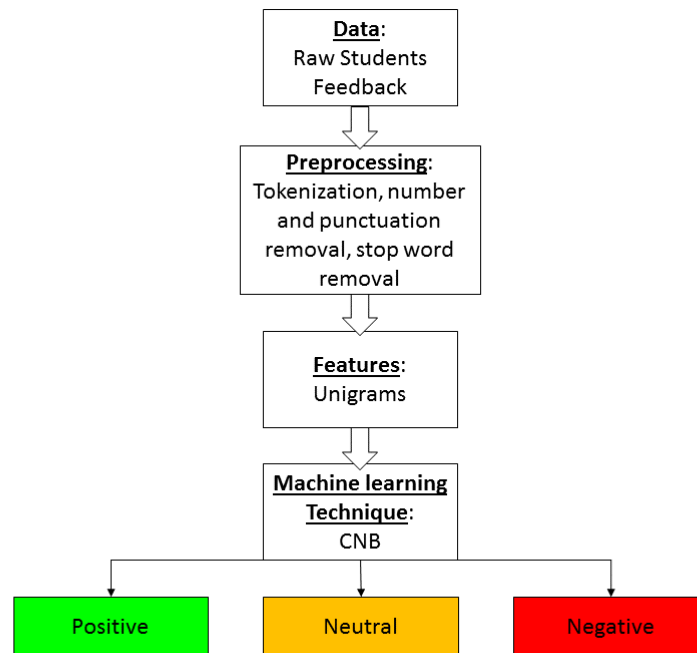


Fig. 4.4 Optimal models process for polarity detection of students' feedback

## 4.5 General Discussion

In this section, we present a discussion of our findings of preprocessing, features and machine learning techniques. We compare between our results and previous findings in the literature of sentiment analysis in education.

Detecting polarity usually includes classifying sentiment into positive, negative and neutral. The results from experiments 1 indicated that preprocessing is important in this study. There are numerous studies that ignore preprocessing, however, our results showed that by preprocessing it increases the accuracy by up to 27% in some models, however in some models the extra preprocessing decreases the accuracy such as in the LIN-Neu model where by adding just stemming, it decreased the results by 46%. In general most of the models performed best with the higher levels of preprocessing from P3-P6.

In experiment 2, we explored a different dataset which was collected via Twitter. Collecting data from Twitter, usually requires a high level of preprocessing due to the noise in the data (i.e. Twitter special characters). We explored 3 different preprocessing levels, and surprisingly from the results we found that the lowest level of preprocessing led to the best performance in most of the classifiers. However, we found that some models perform best by only removing numbers and punctuation. This may indicate that hashtags in the data are valuable to the analysis and contain sentiment. Hashtags can be used for many reasons, however, our results indicate that students use them to express sentiment, therefore they are valuable to polarity detection. Additionally, due to the spelling correction function provided in most phones, it lowers the risk of students making spelling mistakes which usually requires extra preprocessing. We originally believed that punctuations could indicate sentiment, however, our experiments suggest that it does not improve the polarity detection. We also found that emoticons were not useful in our application according to the feature usefulness experiment, which contradicts with previous research in sentiment analysis. This could be due to the students not using it in their feedback (i.e. only 63 instances out of our whole dataset) as it is educational related.

As a conclusion of experiments 1 and 2, we choose a lower level of preprocessing which includes: tokenization, converting text to lower case, removing numbers, punctuation and stop words, which were found to improve the results in both experiments 1 and 2.

As for the features, which are a critical element in sentiment analysis models due to them being the inputs of the classifiers, we experimented with a variety of features in isolation and combination. In experiment 1, we explored the most common features which are n-grams and POS-tagging. The previous literature indicated that both n-grams and POS-tagging features led to a good performance in the educational domain, making it difficult to distinguish which features would be most suitable for our application. We experimented with 12 different combinations of these features. Our results revealed that n-grams features led to a better performance than POS-tagging. We found that the features that led to the best performance were unigrams and unigrams combined with bigrams. Due to unigrams leading to the highest models, we selected it to be experimented with in our second experiment.

In experiment 1, only part of the data was obtained from Twitter, while in experiment 2 the data corpus was all from Twitter. In experiment 2, the aim was to look at less common features in sentiment analysis related to the educational domain and Twitter. We experimented with lexicons, tweet-related, general text-related features and domain-based features. We found that adding these features increased the performance of the models compared with the unigrams feature alone increasing the accuracy by up to 11%. We found that the most valuable and useful features were the lexicons according to the feature usefulness experiment. Additionally, we found some of the features such as emoticons and number of question and exclamation marks were not useful. This could be the reason of their small effect in performance. We found adding 15 features to the unigrams led to the best results. Despite our findings, using these features in real settings may be difficult, due to it requiring extra information from the student, such as the emotion labels.

By observing the results of experiments 1 and 2, we found that these experiments contradict each other. In experiment 1, using unigrams alone and not in combination of other n-grams performed the best. However, in experiment 2, we found that using the extra features in UNI+15 improved the performance of the models.

In conclusion, in sentiment analysis, it is better to use the least number of features to obtain the results due to the less information required. Therefore, by observing the previous results, we chose to use unigrams only. However, these experiments could be useful for future researchers to further explore the importance and the role of features in sentiment analysis models.

The machine learning techniques are the most important element in the sentiment analysis as they produce the models that analyse the data and classifies it. In experiment 1, we experimented with the most common machine learning techniques found in sentiment analysis which are NB, SVM, ME and CNB. We experimented with different kernels in the SVM technique. We found that the polynomial kernel had the lowest performance. This contradicts with previous research that had shown that it performed well in the educational domain. We found that the best kernel was the linear kernel, which led to the highest performing model. We found that classifiers NB, LIN and CNB were the highest performing classifiers across models with the neutral class and without it. Additionally, across all models in the neutral, we found that CNB addresses the problem of the unbalanced datasets.

In experiment 2, we experimented with three additional classifiers that are less common, which are MNB, SMO and RF. We found that MNB performs well with models without the neutral class and SMO performs well with models including the neutral class. We also found in this experiment that CNB is the best classifier for models with the neutral class.

In conclusion to experiments 1 and 2, we found that CNB is the best option due to its high performance with datasets with the neutral class and capability of addressing unbalanced datasets.

Although many studies have excluded the neutral class, we found that including it is important for this study, as students' feedback can be neutral. Excluding the neutral class does not show a complete picture of the class opinion. When the neutral proportion is not known, it can mislead the true proportions of the positive and negative opinions. Our results suggest that by including the neutral class, the performance of the models show only a slight reduction in performance. Moreover, some of the classifiers such as SVM, NB and CNB performed very well including the neutral class. Furthermore, we found in our experiment in chapter 7, that most lecturers when given the choice to include or exclude the neutral class, they preferred to include it. We also found in both experiments that CNB addresses the problem of unbalanced classes, thus our reason for choosing it in our application.

Due to the unbalanced datasets, we explored oversampling and undersampling. We found that undersampling does not have any impact on our data. On the other hand, we found that oversampling increased the performance. However as the CNB addresses the problem of unbalanced classes without oversampling, we did not use it for our application.

This study has several limitations. First, the datasets was small (as already discussed in section 4.3.4) and imbalanced. The neutral class was smaller than the positive and negative. Additionally, in the second experiment, we found that some features occurrences were small in the dataset such as emoticons and number of question and exclamation mark.

## 4.6 Summary and Contribution of the Chapter

In this chapter we explored different models to detect polarity from students' feedback. We performed 2 experiments to explore different features. The first experiment included exploring different preprocessing levels with the most common features which are n-grams and POS-tagging and the most popular machine learning techniques. Additionally in this experiment the neutral class is also explored. We found from this experiment that preprocessing increased the performance of the models. Moreover, the n-gram features led to a better performance than POS-tagging. The best classifiers in this experiment were LIN, NB and CNB. We also discussed the importance of the neutral class in our application and found that CNB performed well with models with the neutral class.

In the second experiment, additional features were explored which were lexicons, tweet-related, general text-related features and domain-based features. We explored with 3 additional classifiers which are MNB, SMO and RF. We also explored with methods to address

unbalanced data. We found that preprocessing the data lowered the performance of the models from our new dataset from Twitter. The preprocessing hashtags were used by students to express sentiment therefore were valuable. The additional features increased the performance of the models compared to unigrams alone. However, we found that some of the features may be difficult to obtain in real-life settings such as emotion labels. We found that undersampling does not have any impact on our data. In contrast, we found that oversampling increased the performance slightly.

In comparison with other domains, we found that using emoticons for the educational domain have little value, despite that the previous literature in other domains have suggested otherwise. In contrast to other studies who eliminated hashtags due to their beliefs that they are not useful, we found them useful in the educational domain.

The contribution of this chapter was to explore polarity detection in students' feedback. One of the main contributions and aims of this chapter was to identify the best performing model for polarity detection in students' feedback. This includes finding the best combination of preprocessing techniques, features and machine learning techniques. The best model found was to use a lower level preprocessing which includes converting text to small case and removing numbers and punctuation, unigrams for features and CNB for the classifier. The other contributions of this chapter include:

- Finding the optimal level of preprocessing. This includes a systematic study of preprocessing levels in sentiment analysis. In the first experiment we found that the higher level of preprocessing increases the performance of the models. On the other hand we found in our second experiment which was from Twitter, that the extra preprocessing decreases the performance of the models.
- Finding what features lead to the best performance for polarity detection in students' feedback. The role of less common features on the models and new features related to our application were explored. We found from the first experiment that n-grams performed better than POS-tagging for our application. In the second experiment we found that by adding other features, the performance of the models increases.
- The creation of a new lexicon based on students' feedback. Lexicon based features were used in the second experiment and increased the performance of the models.
- The exploration of machine learning techniques, which included techniques that have not been experimented with much in sentiment analysis before such as CNB, SMO and RF. We found that the CNB classifier performed well with models with unbalanced classes (i.e. models with the neutral).

- The investigation of the neutral class and its importance and performance in sentiment analysis for students' feedback. Ignoring the neutral class distorts instances into positive and negative classes. Across both our experiments models the neutral class performed well with the CNB classifier.





# Chapter 5

## Detecting emotions from students' feedback

In the previous chapter we presented our experiments on polarity detection and in this chapter we build on the results from these experiments to explore emotion detection. We review previous literature related to emotion detection and explore learning emotions. In this chapter we present two experiments: exploring the n-grams in section 5.2 and exploring other features in section 5.3. This chapter is concluded with the optimal models for emotion detection in students' real-time feedback. Lastly, a summary of the chapter and contribution are presented.

### 5.1 Emotion detection

Emotions are important in the learning process [50, 51, 68]. Positive emotions may increase students' interest in learning, increase engagement in the classroom and motivate students [51]. Additionally, students who are happy are generally more motivated to accomplish their learning goals [118]. On the other hand, emotions such as anger, anxiety and sadness have a negative influence on students and can distract their learning [174].

Emotions can be detected from facial expressions by detecting changes in facial muscles. For instance, Mase [109] detected happiness, anger, disgust and surprise emotions from facial muscles. Gunes and Piccardi [72] research also suggests emotion can be detected from facial expressions. They presented some examples of facial expressions such as in fear emotion, the brows are raised and drawn together and in uncertainty emotion the corners of the lips are drawn downwards.

Emotions can be detected from body gestures. Gunes and Piccardi [72] presented some examples of body gestures suggesting emotion; for instance, when a person's hands are close

to the table surface this reflects anxiety. Additionally, they found that shoulder shrugs reflect uncertainty. One limitation in detecting emotions from body gestures is that they could represent multiple emotions. For example, moving the body backwards suggests both disgust and fear [72].

Emotion can also be detected through speech. Research by Dellaert et al. [45] aimed to classify four emotions with pitch-related features using a machine learning approach. Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR) and K-nearest Neighbours (KNN) classifiers were utilised. They found that these classifiers give an error rate of 32% to 44%. Busso et al. [20] revealed that some words in speech are correlated with specific emotions, which could also be applied when detecting emotions in text. Their study claimed anger accuracy was 95%, but all other emotions accuracies were only 0-2%.

In the educational domain, facial expression and body gestures have been used to detect students' emotions [8, 164]. Arroyo et al. [8] used cameras and sensors to find how students feel and behave while solving mathematics problems in a public school setting. Their facial software helped to predict more than 60% of the students' emotional states. Terzis et al. [164] used the FaceReader tool to detect students' emotions and found it is capable of measuring emotions with an accuracy of over 87%.

Detecting emotions through facial expressions, body gestures and speech can be expensive and complicated due to the equipment (i.e. sensors and cameras) required to detect them. An easier and more efficient method to collect data can be through text. Our research explores emotion detection from text via sentiment analysis.

As mentioned in chapter 4, section 4.1, sentiment analysis has mostly been focused on detection of polarity (negative or positive sentiment) rather than specific emotions. Thus, there is relatively little research on the prediction of specific emotions from text [4, 16, 41, 78, 84, 158, 163, 173, 179, 187]. There are even fewer reports of emotions prediction in education [118, 170]. Moreover, from these studies (both within the educational field and outside of it), an even smaller number use machine learning to predict emotion from text, e.g. [4, 41, 78, 170].

Inkpen [78] classified 234,129 blog posts from Livejournal, a free weblog service used by millions of people to create weblogs into 132 mood classes. They split their data into 144,129 for training data and 90,000 for testing. Four types of features were used: Frequency Counts Bag-of-Words, Length-related Features, Sentiment Orientation and Special Symbols (i.e. emoticons). Chi-square method was used to reduce the number of features in the Frequency Counts Bag-of-Words. SVM was used as a machine learning technique, and their study was evaluated using accuracy, precision, recall and F-score. Their results revealed that the best features were Frequency Counts Bag-of-Words and Sentiment Orientation combined

together, leading to an accuracy of 97.93% in some emotions (i.e. exhausted, restless and exanimate (lifeless)).

Likewise, Aman and Szpakowicz [4] detected six emotions (i.e. happiness, sadness, anger, disgust, surprise and fear) from blog posts. They collected 1466 sentences with emotions and 2800 sentences without it. Three type of features were used: General Inquirer features, WordNet-Affect features and Other features. The General Inquirer features consisted of: Emotion words, Positive words, Negative words, Interjection words, Pleasure words and Pain words. The WordNet-Affect features consisted of: Happiness words, Sadness words, Anger words, Disgust words, Surprise words and Fear words. Other features consisted of: Emoticons Exclamation (“!”) and question (“?”) marks. Naive Bayes and SVM were used as machine learning techniques. They measured their results using accuracy and obtained the best accuracy at 74% with SVM using General Inquirer features and WordNet-Affect features combined.

Dansiman and Alpkocak [41] detected emotions using the International Survey on Emotion Antecedents and Reactions dataset, which consists of 7,666 sentences. They also collected sentences using Wisconsin Perceptual Attribute Rating Database which was collected from undergraduate students using an online form and the Wordnet-Affect dataset. Their research aimed to detect seven emotions: joy, fear, anger, sadness, disgust, shame and guilt. They removed stop words and identified question and exclamation marks. They used Vector Space Model (VSM), ConceptNet, Naive Bayes and SVM as classification techniques. They found the VSM classifier gave the highest performance with an overall accuracy of 67% and F-measure value of 32.22% across all emotions. This data, however, is not representative for other types of text expressing emotions, as indicated by the low accuracy, i.e. less than 35%, of these models on test sets with other data.

Yuan and Purver [187] investigated the prediction of emotions (i.e. happiness, sadness, anger, disgust, surprise and fear) from the Chinese microblog service Sina Weibo. Their database contained 229,062 Weibo statuses with their corresponding emotion labels. To preprocess the data, they removed URLs, digits and other notations. Different kinds of lexical features were used: segmented Chinese words, Chinese characters, and n-grams (unigrams, bigrams and trigrams). SVM was used as a classifier and three emotion accuracies were reported angry, happiness and sadness giving 70.1%, 78.2% and 69.6%, respectively.

Strapparava and Mihalcea [158] study detected six emotions: Anger, Disgust, Fear, Joy, Sadness and Surprise, from 1250 news headlines. They used three methods to detect emotions SWAT, UA, UPAR7. SWAT is a supervised system that uses unigram features in a model trained to annotate emotional content. The UA method consisted of using POS-tagging as features (i.e. nouns, verbs, adverbs and adjectives) and Pointwise Mutual Information to

count emotion scores. The UPAR7 method consisted of parsing the headlines through a Stanford syntactic parser, then emotions were detected using a lexicon-based approach with SentiWordNet and WordNetAffect lexicons. SentiWordNet is a lexicon consisting of 17,530 terms classified into objective, positive and negative. WordNetAffect is a lexicon containing 1,903 terms directly or indirectly referring to mental (e.g. emotional) states. Strapparava and Mihalcea [158] found in general that the UPAR7 method led to the best accuracy across most of the emotions. However, the highest accuracy score was obtained with the UA method for the disgust emotion at 97%.

Teng et al. [163] collected 1200 Chinese sentences for each emotion (i.e. neutral, happiness, anger and sadness). Their dataset consisted of 960 sentences used as a training set and 240 used as a testing set. Keywords were used as features and the Rough Set Theory was used to find a minimal subset from the attribute set with SVM. Consequently, by using the Rough Set Theory the predicted accuracies of their models were improved in comparison to without using it. The sum of averages from all the emotion classes was 87.5%.

Tromp and Pechenizkiy [173] detected eight emotions derived from Plutchik's model: joy, sadness, trust, disgust, fear, anger, surprise and anticipation. They used two datasets: The Affect dataset which includes 12689 instances of sentences and the Twitter dataset which includes 1084 sentences from three languages (i.e. English, Dutch and German). Two thirds of their data were used as a training set and one third as a testing set. They also experimented the neutral class in both the datasets. The authors initiated a new classification algorithm 'RBEM-Emo' and compared its performance against other classifiers Support Vector Machines (SVMs), regression and the recursive auto-encoder [155]. They found that their classifier performed better than other classifiers such as SVM. In addition, the highest accuracy occurred in the Affect dataset with an accuracy of 88%.

Binali et al. [16] detected 22 emotions from the OCC (Ortony/Clore/Collins) model. Their dataset consisted of 2340 of training instances and 540 of testing instances. Their model consisted of two classes: emotion and non-emotion. A hybrid approach was used consisting of a combination of keyword based implementation and learning based implementation. Keyword based implementation relies on the presence of keywords in an emotion lexicon. Learning based implementation uses a trained classifier to categorise input text into emotion classes by using keywords as features. They used SVM as a classifier which gave a 96% accuracy.

Keshtkar and Inkpen [84] extracted paraphrases for emotions from texts. Six emotions were detected: happiness, sadness, anger, disgust, surprise and fear. They collected 24299 sentences from different blog datasets. Bootstrapping Algorithm was used to extract paraphrases and human judges evaluated the correctness. The authors used precision and recall

and consequently, their results included an average of 84% precision and 89% recall across all emotions.

Chaffar and Inkpen [26] used multiple datasets from blogs to detect 6 emotions: anger, disgust, fear, happiness, sadness and surprise. The dataset was collected from newspapers and the Google News search engine. They used 1,000 annotated sentences for testing and 250 annotated sentences for training. Three types of features were used: Bag of Words, n-grams and lexicon emotion features (i.e. using WordNetAffect). They used a machine learning approach using three classifiers: Naive Bayes, J48 and SMO. The authors found that SMO performed best with an accuracy rate of 81%.

Wang et al. [179] study detected 8 emotions (expect, joy, love, surprise, anxiety, sorrow, hate and anger) from a Chinese dataset. Their dataset consisted of 1487 Chinese blogs and 47740 emotional sentences. Four features were used in their experiments: Keyword, Keywords and POS-tagging, Keywords and Intensity and Keywords, POS-tagging and Intensity combined. They used ME as a classifier and found that love and surprise emotions led to the highest accuracy at 90% with all the features combined. This study also explored a multi-class model with all the emotions combined which resulted in a low accuracy of 35% with the Keywords and Intensity features.

As for the educational domain, Tian et al. [170] predicted students' emotions from text [170]. They proposed a system to recognize Chinese e-learner's emotions based on interactive texts. Similarly, Munezero et al. [118] predicted students' emotions through their learning diaries.

The process to detect emotions from text is the same as the polarity detection process mentioned earlier in chapter 4, in section 4.1. An overview of research related to the first three aspects, i.e. preprocessing, features and machine learning techniques, is given in the following.

### 5.1.1 Preprocessing

Preprocessing the text for emotions detection is important as in polarity detection. Similar to the polarity experiments in chapter 4 in section 4.1.1 both general preprocessing techniques and Twitter-related preprocessing techniques have been used for emotion detection. For instance, Danisman and Alpkocak [41] study used several preprocessing techniques such as removal of stop words, stemming and negation handling. They also replaced exclamation marks and question marks with "XXEXCLMARK" and "XXQUESMARK". Keshtkar and Inkpen [84] used tokenization and removed URLs from their data which was obtained from blogs. Most of the studies in emotion detection have not used preprocessing or only used low level preprocessing such as tokenization only [4, 16, 78, 158, 163, 173].

### 5.1.2 Features

Similar to polarity detection in chapter 4, n-grams are the most popular features in emotion detection [158]. The most commonly used n-gram is unigrams, i.e. single words, as found in many research works in emotion detection, e.g. [158] and [78]. In contrast, there are very few studies investigating the use of bigrams and trigrams in emotion prediction. Consequently, in section 5.2, we investigate the influence of these different n-grams and their combination on emotion detection.

Other features that can be used in emotion detection include lexicon-derived features, number of hashtags and emoticons, number of punctuation signs and question and exclamation marks. In section 5.2, we experiment with these features.

Qadir and Riloff [137] used lexicons in combination with machine learning techniques. They used the NRC Emotional Tweets Lexicon to label four emotions from tweets; anger, fear, joy and sadness. The NRC lexicon contains 14,182 unigrams and emotion labels (i.e. anger, anticipation, disgust, fear, joy, sadness, surprise and trust). Qadir and Riloff [137] added a hashtag to words to match other hashtags and their emotion in the lexicon. They found that hashtags such as ‘#anger’ holds sentiment. Hashtags are one of Twitter related features, which can be used to mark topics [1]. They can also be used to express emotion [137], therefore they may be a valuable feature in the detection of emotion. Similarly, Wang et al. [181] found hashtags useful when labelling tweets. They also used lexicons LIWC Lexicon, MPQA Lexicon and WordNetAffect to detect 8 emotions (i.e. love, joy, surprise, anger, sadness, fear, surprise and thankfulness). The LIWC lexicon contains over 4000 English words and stems and labels data into positive and negative emotions (contains subdictionaries for Anger, Anxiety and Sadness) [7]. The MPQA affect lexicon contains 11021 instances of words with emotion and without [30].

Emoticons is another feature that may be valuable in emotion detection. Emoticons are widely used in social networks like Twitter [47]. They can be used to determine the polarity or emotion label for a sentence. For instance, smiley emoticons can reflect happy polarity or emotion [1]. Emoticons are popular in sentiment analysis studies [1, 47, 78, 92, 136, 190]. Agarwal et al. [1] used emoticons to determine the polarity of the sentence. Dhawan et al. [47] used the number of emoticons as a feature to detect emotions (i.e. happy, sad, angry, disgust, fear, surprise) from text.

Number of punctuation signs and number of question and exclamation marks could be an important feature when detecting emotion in text, as question marks may mean confusion and exclamation marks may mean strong emotion [190]. Knowing the exact number of punctuations and question/marks is irrelevant, however, knowing whether the person uses

few or many punctuation marks might indicate some sort of subjectivity [184]. Dhawan et al. [47] used the number of punctuation marks as features.

### 5.1.3 Machine Learning Techniques

Various *machine learning techniques* have been used to predict emotion from text. The classifiers that have been previously experimented with for emotion prediction are: Naive Bayes (NB) [4, 41, 173], Multinomial Naive Bayes (MNB) [170], Complement Naive Bayes (CNB) [170], Support Vector Machines (SVM) [4, 41, 163, 173], Maximum Entropy (ME) [179], Sequential Minimal Optimization (SMO) [26] and Random Forest (RF) [170].

### 5.1.4 Learning Emotions

Previous research on emotions related to learning indicates a variety of emotions. An overview of this research is given in Table 5.1. Some studies grouped similar emotions, which may be beneficial as using too many emotions can be difficult to classify and may cause conflicts [89, 100]. This, however, could be misleading for the real emotion, such as in the case of grouping disgust with boredom or frustration with sadness. For example, Tian et al. [169] study about e-learners emotions grouped “disgust and boredom” and “frustration and sad”. They also included different levels and strengths for each emotion such as weak and strong, which may be confusing for e-learners to determine and difficult for the classifier to determine. Similarly, Kort et al. [89] grouped opposite emotions and their strengths. Some of the studies missed important emotions; for instance O'Regan [122] missed boredom and confusion emotions which are important in learning.

We identified 19 common emotions that are associated with learning: amused, anxiety, appreciated, awkward, bored, confused, disappointed, embarrassed, engaged, enthused, excited, frustrated, happy, motivated, proud, relief, satisfied, shame and uninterested. These emotions were chosen due to the frequency in previous studies and relevance to learning.

## 5.2 Emotion Detection Models for Students' Feedback: N-gram features

In this section, we explore emotion detection using different preprocessing levels and n-gram combinations as features. We present the data corpus and the experiment, then discuss the results.

Table 5.1 Summary of studies with emotions

Research	Emotions
D'Mello and Graesser [50]	Confusion, Interest, Frustration, Excitement, Boredom, Insight
Nosu and Kurokawa [120]	Easy, Difficult, Boring, Interesting, Confused, Comprehending, Tired, Concentrating
O'Regan [122]	Frustration, Pride, Fear, Anxiety, Apprehension, Shame / Embarrassment, Enthusiasm / Excitement
Shen et al [153]	Interest, Boredom, Engagement, Hopelessness, Evaluation, Satisfaction, Frustration, Disappointment
Kort et al. [89]	(Anxiety-Confidence): Anxiety, Worry, Discomfort, Comfort, Hopeful, Confident (Boredom-Fascination): Ennui, Boredom, Indifference, Interest, Curiosity, Intrigue (Frustration-Euphoria): Frustration, Puzzlement, Confusion, Insight, Enlightenment, Ehipany (Dispirited-Encouraged): Dispirited, Disappointed, Dissatisfied, Satisfied, Thrilled, Enthusiastic (Terror-Enchantment): Terror, Dread, Apprehension, Calm, Anticipatory, Excited
Litman and Forbes [100]	Negative: Uncertain, frustration, boredom, confusion Positive: Confident, Interested, engaged, encouraged Neutral: No strong expression of the emotion
Tian et al. [169]	Appreciation (AP), Anger (AN), Sympathism (SY), Inhospitallity (IN), Love/adore (LO), Disgust/boredom (BO), Relief (RE), Anxiety (AX), Joy (JO), Frustration/sad (SA), Pride (PR), Shame and guilt (SG), Hope (HO), Hopeless (HL), Positive surprise (PS)
Tian et al. [168]	Love, Joy, Anger, Frustration, Neutral



### 5.2.1 Data Corpus

The data is from the same source as the data collected in the previous chapter 4. It was collected from lectures taught in English in Jordanian universities on different topics: calculus, English communication skills, database, engineering, molecular biology, chemistry, physics, science, contemporary history of the world and architecture.

As previously mentioned in chapter 4, Twitter was used to collect students' feedback about the lecture. For each tweet, they were asked to choose one emotion from a set of emotions provided, i.e. the 19 emotions listed in the previous section.

A total number of 1522 tweets were collected with their corresponding emotion label. Some of the emotions appeared more frequently than others. The most frequent emotions presented in Table 5.2 were used in our experiments, while the least frequent ones presented in Table 5.3 were discarded due to insufficient data for training and testing machine learning models.

Table 5.2 Most Frequent Emotions

Most Frequent	No. of feedback
Bored	336
Amused	216
Frustrated	213
Excited	178
Enthused	176
Anxiety	130
Confused	73
Engaged	67
<b>Total used</b>	<b>1389</b>

Table 5.3 Less Frequent Emotions

Less Frequent	No. of feedback
Happy	32
Satisfied	31
Appreciation	26
Embarrassed	18
Dissapointed	12
Uninterested	4
Proud	3
Relief	3
Shame	2
Awkward	1
Motivated	1
<b>Discarded</b>	<b>133</b>

### 5.2.2 Experiments

Due to our findings in the previous literature and our findings from the previous chapter where the best results were either with a high level or a low level preprocessing, we experimented with two different preprocessing levels: (a) high preprocessing, which includes: tokenization, covert text to lower case, remove punctuation, remove numbers, remove stop words, remove hashtags, remove URLs, remove retweets, remove user mentions in tweets and remove

Table 5.4 Abbreviations and number of features

Features	# of Features
Unigrams (UNI)	2989
Bigrams (BI)	11425
Trigrams (TRI)	14314
Unigrams and Bigrams combined (UNI+BI)	17317
Unigrams and Trigrams combined (UNI+TRI)	19654
Bigrams and Trigrams combined (BI+TRI)	26608
Unigrams, Bigrams and Trigrams combined (UNI+BI+TRI)	28720

Twitter special characters; (b) low processing, which includes: tokenization, covert text to lower case and remove stop words.

The high preprocessing was only used for one of the models which contained all the emotions combined. The results showed that for this model, the lower level of preprocessing performed better than the higher level. Consequently, for the other models, only the low level of preprocessing was explored.

We experimented with different n-grams, i.e. unigrams, bigrams and trigrams, and every possibility in combining them to find which n-gram or combination of n-grams leads to the best performance for the different models. The features that were experimented with are presented in Table 5.4.

We used Naive Bayes (NB) and Support Vector Machines (SVM) with two common kernels: radial basis (RB) and linear (LIN) kernels due to their common use in previous research. Additionally we use less common machine learning techniques due to their promising results: Multinomial Naive Bayes (MNB), Maximum Entropy (ME), Complement Naive Bayes (CNB) Sequential Minimal Optimization (SMO) and Random Forest (RF).

We experimented with all the emotions combined and then subtracted, in turn, the emotion with the lowest number of data. The total number of models experimented with was 16 models, which are illustrated in Table 5.5.

### 5.2.3 Results and Discussion

All the models were tested using 10-fold cross-validation; the accuracy and the error rate were used to assess the overall performance of the classifiers, while the true positive rate (i.e. recall) and precision were used to assess the ability of the classifiers to correctly identify the specific emotion(s).

The highest accuracy results for all models are displayed in Table 5.6, while the highest True Positives (TP) results for all models are given in Table 5.7. In both tables the TP rate (i.e.

Table 5.5 The classes and preprocessing levels used for the models

Model (Emotions)	Pre-processing level	No. of classes
All emotions	High	8-Classes
All emotions	Low	8-Classes
7 emotions (Amused, Anxiety, Bored, Confused, Enthused, Excited, Frustrated) +other	Low	8-Classes
6 emotions (Amused, Anxiety, Bored, Enthused, Excited, Frustrated) + other	Low	7 classes
5 emotions (Amused, Bored, Enthused, Excited, Frustrated) + other	Low	6 classes
4 emotions (Amused, Bored, Excited, Frustrated) +other	Low	5 classes
3 emotions (Amused, Bored, Frustrated) +other	Low	4 classes
2 Emotions (Amused, Bored) + other	Low	3-Classes
Each emotion + other (8 models)	Low	2-Classes

recall) and precision are given for the emotion(s) class(es). For the multi-emotion models the reported values are the average of the emotion classes (i.e. excluding the 'other' class).

The results indicate that the models with a single emotion perform better than the multi-emotion models in terms of accuracy, although one has to bare in mind that the baseline for multi-class models is lower than the baseline for 2-Class models.

Table 5.6 shows that two classifiers performed best in terms of accuracy: the Support Vector Machine with Radial Basis kernel (RB), mainly for the 2-Class models, and Sequential Minimal Optimization (SMO), mainly for the multi-class models. In term of features, unigrams and trigrams were found to lead to the best performance for the 2-Class models, while unigrams combined with bigrams and trigrams led to the best performance for the multi-class models.

Despite the fact that accuracy can be useful in predicting the model's performance, it does not indicate how well a classifier can predict specific emotions. As the true positive rate (or recall) indicates the percentage of correctly identified instances for a class of interest, it can be used to assess the ability of the classifiers to predict emotions; in addition, precision can indicate where the identification problems occur.

For most of the models with the highest accuracy, the true positives rate is extremely low or even 0% in some cases. In addition, precision is also low (with a few exceptions). This indicates that the high accuracy is due to the correct identification of the 'other' class rather than the correct identification of emotion(s).

Table 5.6 Highest accuracy for each model

Dataset	Technique	N-gram	Accuracy	Precision	Recall	F-score	Error rate
ALL Preprocessed	CNB	UNI+BI+TRI	0.32	0.32	0.34	0.33	0.68
ALL W/O Preprocessing	CNB	UNI+BI+TRI	0.34	0.31	0.32	0.31	0.66
7 Emotions+ other	SMO	UNI+BI+TRI	0.28	0.20	0.21	0.20	0.72
6 Emotions+ other	SMO	UNI+BI	0.29	0.25	0.26	0.25	0.71
5 Emotions+ other	SMO	UNI+TRI	0.31	0.30	0.24	0.27	0.00
4 Emotions+ other	SMO	UNI+BI	0.38	0.41	0.13	0.20	0.62
3 Emotions + other	ME	UNI+TRI	0.51	0.47	0.29	0.36	0.49
2 Emotions+ other	RB	TRI	0.61	0.35	0.03	0.06	0.39
Amused	RB	UNI	0.84	0.00	0.00	0.00	0.16
Anxiety	RB	UNI	0.91	0.00	0.00	0.00	0.09
Bored	RB	TRI	0.76	0.30	0.01	0.02	0.24
Confused	RB	UNI	0.95	0.00	0.00	0.00	0.05
Engaged	RB	UNI	0.95	0.00	0.00	0.00	0.05
Enthused	RB	UNI	0.87	0.00	0.00	0.00	0.13
Excited	RB	UNI	0.87	0.00	0.00	0.00	0.13
Frustrated	SMO	TRI	0.85	0.57	0.09	0.16	0.15

Table 5.7 Highest True Positives for each model

Dataset	Technique	N-gram	Accuracy	Precision	Recall	F-score	Error rate
ALL Preprocessed	ME	UNI+BI+TRI	0.32	0.34	0.33	0.33	0.67
ALL W/O Preprocessing	ME	UNI+BI	0.32	0.33	0.32	0.32	0.68
7 Emotions+ other	NB	BI+TRI	0.26	0.24	0.25	0.25	0.75
6 Emotions+ other	MNB	UNI	0.27	0.27	0.26	0.27	0.73
5 Emotions+ other	MNB	UNI+TRI	0.25	0.32	0.32	0.32	0.68
4 Emotions+ other	MNB	BI	0.26	0.29	0.38	0.33	0.67
3 Emotions + other	ME	UNI+BI+TRI	0.51	0.43	0.36	0.39	0.61
2 Emotions+ other	ME	UNI+BI+TRI	0.57	0.40	0.51	0.45	0.55
Amused	CNB	TRI	0.49	0.19	0.70	0.30	0.70
Anxiety	CNB	TRI	0.45	0.12	0.77	0.21	0.79
Bored	CNB	TRI	0.44	0.28	0.85	0.42	0.58
Confused	CNB	TRI	0.28	0.06	0.81	0.11	0.89
Engaged	CNB	TRI	0.24	0.04	0.68	0.08	0.92
Enthused	CNB	TRI	0.36	0.14	0.76	0.24	0.76
Excited	CNB	TRI	0.37	0.15	0.86	0.26	0.74
Frustrated	CNB	TRI	0.40	0.19	0.84	0.31	0.69

Table 5.8 Best overall models for identification of specific emotions

Dataset	Technique	N-gram	Accuracy	Error rate	Precision	Recall	F-score
Amused	CNB	Bi	0.63	0.37	0.24	0.60	0.34
Amused	CNB	Bi+Tri	0.64	0.36	0.24	0.62	0.35
Bored	MNB	UNI	0.70	0.30	0.41	0.53	0.46
Bored	CNB	UNI	0.66	0.34	0.38	0.63	0.47
Bored	CNB	BI	0.63	0.37	0.36	0.67	0.47
Bored	MNB	UNI+BI	0.73	0.27	0.44	0.55	0.49
Bored	CNB	UNI+BI	0.68	0.32	0.39	0.63	0.48
Bored	CNB	UNI+TRI	0.55	0.45	0.31	0.71	0.43
Bored	MNB	BI+TRI	0.73	0.27	0.44	0.55	0.49
Bored	CNB	BI+TRI	0.68	0.32	0.39	0.63	0.48
Bored	ME	BI+TRI	0.73	0.27	0.45	0.54	0.49
Bored	MNB	UNI+BI+TRI	0.73	0.27	0.45	0.58	0.51
Bored	CNB	UNI+BI+TRI	0.71	0.29	0.43	0.63	0.51
Excitement	CNB	BI	0.63	0.37	0.20	0.61	0.30
Excitement	CNB	UNI+TRI	0.64	0.36	0.21	0.64	0.32

In contrast to Table 5.6, Table 5.7 displays the best experimental results when focusing on the true positives rate, i.e. the correct identification of the emotion(s). In terms of machine learning techniques, Complement Naive Bayes (CNB) performs best for half of the models, which could be explained by the ability of this technique to compensate for uneven class sizes. In terms of features, trigrams led to the best performance in the 2-Class models, while unigrams combined with bigrams and trigrams led to the best performance in the multi-class models.

The fact that the models with high true positives rates have low accuracy and low precision values indicates that many instances of the 'other' class are wrongly classified as particular emotions. In other words, although the classifiers have a higher sensitivity for the emotion classes, they are not precise in distinguishing the 'other' class from the emotion class(es).

When looking at the overall picture and the balance of the evaluation metrics considered (i.e. accuracy, true positive rate, error rate and precision), some of the models stand out – these are presented in Table 5.8. We found that the best classifiers are Complement Naive Bayes (CNB) and Multinomial Naive Bayes (MNB). When observing at the features, one can notice that for only 2 out of 15 models, unigrams led to the best performance, while the other 13 models used other n-grams or combinations of n-grams. This indicates that bigrams and trigrams, as well as combinations of various n-grams are useful for the prediction of specific emotions and should be investigated further.

It is not surprising that the best performing models are for the emotions for which we had larger number of instances (see Table 5.2 in section 5.2.2), i.e. bored, amused and excited. Interestingly, the models for excited performed better than the ones for frustration, although there were more instances for frustration than excited.

From previous research studies focusing on the prediction of emotions using machine learning techniques, only one study was conducted in an educational context [170]. This research used part-of-speech (POS) tags as features, and more specifically, they experimented with the combination of the following part-of-speech tags: verb, adverb, adjective and noun. Their models were evaluated using precision, recall and F-score and consequently their results revealed that Random Forest performed better than the other classifiers with a weighted average F-score at 0.638. Similar to our research, they found that the recall score was higher than the precision. From the emotions that we identified as relevant for learning from previous literature, they only looked at anxiety, for which they obtained a precision value of 0.6 using a LogitBoost classifier. However, this research was conducted on Chinese text, which has different characteristics and structures compared with English text. Moreover, the research was based on text from online chats and discussion groups. Furthermore, they used in their approach an affective words base (i.e. lexicon), where each affective word had a number associated with its degree of reflection of a particular emotion.

There are several studies outside the domain of education exploring the use of unigrams, however, very few studies investigating the use of other n-grams. Youn and Purver [187] found in their research on prediction of emotions from the Chinese microblog that models with bigrams and trigrams outperformed models using unigrams. Similarly, our results showed that using all of the n-grams (i.e. unigrams, bigrams and trigrams) combined led to the best identification of emotions for the multi-emotion models. Additionally, we found that trigrams led to the best identification of emotions for the 2-class models.

In contrast to our study, Aman [4] did not discuss their results in terms of identifying the presence of emotion (i.e. recall rate). They presented the accuracies of the models, which is misleading due to the presence of classes without emotion.

### **5.3 Emotion Detection Models for Students' Feedback: Other features**

In this section, we explore emotion detection using less common features including features related to the educational domain and Twitter. The purpose of this study is to build on to the

first experiment and explore the role of features. We present the data corpus first and then discuss the experiment and results.

### 5.3.1 Data Corpus

The datasets for this experiment were the same as the ones in section 5.2. In the first dataset, we chose to experiment with all the emotions to investigate if the extra features increase the results of the previous experiment. The rest of the datasets were chosen due to their high performance in section 5.2. The datasets are as follows:

- 8-Class dataset which contains labels of all of the emotions listed in 5.2.1;
- 3-Class dataset which contains three classes (i.e. Amused, Bored, Other);
- 2-Class Amused dataset which contains two classes (i.e. Amused and Other);
- 2-Class Bored dataset which contains two classes (i.e. Bored and Other); and
- 2-Class Excited dataset which contains two classes (i.e. Excited and Other).

### 5.3.2 Experiments

Due to the low results associated with preprocessing in experiment 1 in section 5.2, only basic preprocessing techniques were used which are tokenization and converting the text to lower case.

In addition to the unigrams used in the previous experiment, we experimented with several other features, which were mentioned earlier in section 5.1.2. In addition to these features, we also explore the time of the lecture and sarcasm labels as features, which have not been explored before in emotion detection but could be useful in our application. We classify these emotions in four categories:

1. Domain-based features:
  - Lexicon-based features (L), i.e. positive, negative, neutral
  - Time of the tweet during the lecture (T), i.e. beginning (B), middle (M), end (E).
2. Sarcasm (S), i.e. sarcastic and non-sarcastic labels
3. Twitter related features
  - Number of emoticons (NE)

- Number of hashtags (H)

#### 4. Punctuation-related features

- Number of all punctuation characters (PC) (e.g. commas, semicolons, brackets, apostrophes, etc.)
- Number of question and exclamation marks (QE)

All combinations of these additional features with unigrams were experimented with. We converted the nominal features to numeric in Weka to find each out the importance of each single feature. For example, the Polarity feature that was originally nominal was converted into several numeric features (positive, negative and neutral lexicon-based features). Each lexicon-based feature was assigned a value 1 if the sentence was labelled by that polarity label and 0 if it did not. The lexicon created in chapter 4 in section 4.3.2 was used in this experiment.

We used the following machine learning techniques: Naive Bayes (NB), Multinomial Naive Bayes (MNB), Complement Naive Bayes (CNB), Maximum Entropy (ME), Support Vector Machines (SVM), Sequential Minimal Optimization (SMO) and Random Forest (RF), due to their popularity and promising results in previous research and in chapter 4.

### Feature Usefulness

Feature usefulness was calculated using 5 statistical tests Chi Square, Information Gain, Gain Ratio, Correlation and OneR which were discussed in chapter 4 in section 4.3.2.

The results of the feature usefulness test is presented in Appendix C. We found that across all the tests there was a similarity in the order of the features according to their usefulness. In general, most of the features were at the same level in feature usefulness or just one level place difference. Similar to the polarity experiment of other features in section 4.3.2, OneR test was the most different in comparison to other feature usefulness tests. As we previously mentioned in section 4.3.2, this could be due to its functionality in using the minimum-error attribute for prediction. We chose Chi Square feature usefulness, as the results were relatively similar across tests. The features we experimented with according to the Chi square method are presented in Table 5.9.

We experimented with all the features, then for each dataset we took the top half of the features (i.e. 6 features) and tested them incrementally to investigate their individual performances. We also experimented with lexicons as a feature, due to their popularity in research. Moreover, we explored with the sarcasm label as a feature. The total number of experiments for each dataset are 8, which are presented in Table 5.10



Table 5.9 Feature usefulness

<b>8-Class</b>	<b>3-Class</b>	<b>2-Class-Amused</b>	<b>2-Class-Bored</b>	<b>2-Class-Excited</b>
Positive lexicon	Positive lexicon	Positive lexicon	Positive lexicon	Positive lexicon
Negative lexicon	Negative lexicon	Negative lexicon	Negative lexicon	Negative lexicon
No. of hashtags	No. of hashtags	Neutral lexicon	No. of hashtags	No. of hashtags
No. of punctuation	Sarcasm	No. of hashtags	Sarcasm	Time-End
Neutral lexicon	No. of punctuation	No. of punctuation	No. of punctuation	Sarcasm
No. of Emoticons	Neutral lexicon	Sarcasm	Neutral lexicon	Neutral lexicon
No. of Question and Exclamation marks	No. of Question and Exclamation marks	No. of Question and Exclamation marks	No. of Question and Exclamation marks	Time-Middle
Time-Beginning	No. of Emoticons	No. of Emoticons	No. of Emoticons	Time-Beginning
Sarcasm	Time-Beginning	Time-Beginning	Time-Beginning	No. of Emoticons
Time-End	Time-Middle	Time-Middle	Time-Middle	No. of punctuation
Time-Middle	Time-End	Time-End	Time-End	No. of Question and Exclamation marks

Table 5.10 Experiments

Setup	8-Class	3-Class	2-Class- Amused	2-Class- Bored	2-Class- Excited	Addit- ional Fea- tures
Setup A	Positive and Negative Polarity	Positive and Negative Polarity	Positive and Negative Polarity	Positive and Negative Polarity	Positive and Negative Polarity	2
Setup B	Positive and Negative Polarity, No. of Hashtags	Positive and Negative Polarity, No. of Hashtags	Positive, Negative and Neutral Polarity	Positive and Negative Polarity, Sarcasm	Positive and Negative Polarity, No. of Hashtags	3
Setup C	Positive and Negative Polarity, No. of Hashtags, No. of Punctuation	Positive and Negative Polarity, No. of Hashtags, Sarcasm	Positive, Negative and Neutral Polarity, No. of Hashtags	Positive and Negative Polarity, Sarcasm, No. of Hashtags)	Positive and Negative Polarity, No. of Hashtags, Time-End	4
Setup D	Positive and Negative Polarity, No. of Hash- tags, No. of Punctuation, Neutral	Positive and Negative Polarity, No. of Hash- tags, No. of Punctuation, Sarcasm, No. of Punctuation	Positive, Negative and Neutral Polarity, No. of Hash- tags, No. of Punctuation	Positive and Negative Polarity, Sarcasm, No. of Hash- tags, No. of Punctuation)	Positive and Negative Polarity, No. of Hash- tags, Time-End, Sarcasm	5
Setup E	Positive and Negative Polarity, No. of Hash- tags, No. of Punctuation, Neutral, No. of Emotion	Positive and Negative Polarity, No. of Hash- tags, Sarcasm, No. of Punctuation, Neutral	Positive, Negative and Neutral Polarity, No. of Hash- tags, No. of Punctuation, Sarcasm	Positive and Negative Polarity, Sarcasm, No. of Hash- tags, No. of Punctuation, Neutral	Positive and Negative Polarity, No. of Hash- tags, Time-End, Sarcasm, Neutral	6
Setup F	Polarity (All), Sarcasm, No. of Hashtags, No. of Punctu- ation, No. of question and exclamation marks, No. of Emoticons, Time (All)	Polarity (All), Sarcasm, No. of Hashtags, No. of Punctu- ation, No. of question and exclamation marks, No. of Emoticons, Time (All)	Polarity (All), Sarcasm, No. of Hashtags, No. of Punctu- ation, No. of question and exclamation marks, No. of Emoticons, Time (All)	Polarity (All), Sarcasm, No. of Hashtags, No. of Punctu- ation, No. of question and exclamation marks, No. of Emoticons, Time (All)	Polarity (All), Sarcasm, No. of Hashtags, No. of Punctu- ation, No. of question and exclamation marks, No. of Emoticons, Time (All)	11
Setup G	Polarity (All)	Polarity (All)	Polarity (All)	Polarity (All)	Polarity (All)	3
Setup H	Sarcasm	Sarcasm	Sarcasm	Sarcasm	Sarcasm	1

### 5.3.3 Results and Discussion

All the models were tested using 10-fold cross-validation; accuracy, error rate, precision, recall, F-score and Area Under Curve (AUC) were calculated. In a Receiver operating characteristic (ROC) curve, the true positive rate is plotted in a function of the false positive rate for different cut-off points of a parameter. The points on the ROC curve represent sensitivity/specificity pairs corresponding to various decision thresholds. The area under the ROC curve (AUC) is a measure of how well a parameter can distinguish between two groups (emotion/not emotion).

The best performing models leading to good performance results – with the corresponding preprocessing levels, features and machine learning techniques are displayed in Table 5.11. These models were chosen due to the highest identification of emotion (recall). The baseline feature was unigrams, we displayed the unigram results with the same machine learning technique as Table 5.11 in Table 5.12. While these results are discussed in more detail in the subsections below, they also convey several high level observations as in the following:

- The accuracy of the models with unigrams alone (see Table 5.12) increased by 11% with the new features, however in some models, the accuracy decreased 12%;
- The emotion detection (recall) increased by 20% in some models;
- CNB performed best in 3 out of 5 datasets; and
- The features (Experiments) led to different performance in each model;

Table 5.11 Best models

Model	8-Class	3-Class	2-Class- Amused	2-Class- Bored	2-Class- Excited
Experiment	Setup E	Setup C	Setup B	Setup F	Setup E
Machine Learning Technique	NB	ME	CNB	CNB	CNB
Accuracy	0.36	0.51	0.70	0.68	0.55
Error rate	0.64	0.49	0.30	0.32	0.45
Precision	0.37	0.35	0.27	0.39	0.20
Recall	0.36	0.57	0.57	0.72	0.80
F-score	0.36	0.43	0.37	0.46	0.31
AUC	0.64	0.72	0.62	0.58	0.66

Table 5.12 Best models (unigrams)

Model	8-Class	3-Class	2-Class- Amused	2-Class- Bored	2-Class- Excited
Feature	unigrams	unigrams	unigrams	unigrams	unigrams
Machine Learning Technique	NB	ME	CNB	CNB	CNB
Accuracy	0.25	0.57	0.65	0.66	0.67
Error rate	0.75	0.43	0.35	0.34	0.33
Precision	0.27	0.39	0.23	0.38	0.21
Recall	0.18	0.45	0.57	0.63	0.60
F-score	0.21	0.42	0.33	0.47	0.32
AUC	0.59	0.67	0.61	0.65	0.64

Table 5.13 Difference between best models in Setups A-H and unigram models

Model	8-Class	3-Class	2-Class- Amused	2-Class- Bored	2-Class- Excited
Setup	Setup E	Setup C	Setup B	Setup F	Setup E
Accuracy%	11%	-6%	5%	2%	-12%
Error rate%	-11%	6%	-5%	-2%	12%
Precision %	10%	-4%	4%	1%	-1%
Recall%	18%	12%	1%	9%	20%
F-score%	8%	1%	4%	-1%	-1%
AUC	5%	5%	1%	-7%	2%

### Feature results

Various experiments from section 5.3.2 led to an increase performance of the unigrams feature models. The difference between the models with unigrams and the highest models in Table 5.11 are presented in Table 5.13.

Table 5.14 presents a comparison between the recall for the unigrams and the Setups with the 8-Class dataset. In the table we have highlighted the best and worst performance in the Setups. The best feature combination in the 8-Class dataset with all emotions, was Setup E which were the highest 6 features (i.e. positive-lexicon, negative-lexicon, number of hashtags, number of punctuation, neutral polarity and number of emoticons). The Setups all led to an increase in the identification of emotions in comparison with unigrams alone, except for Setup A for the CNB classifier where there was a decrease in the recall by 2%. We summed the increase in the accuracy and subtracted the decrease for each of the setups to find the least effective Setup. Consequently, we found that the lowest increase happened in Setup H which was just the sarcasm feature alone.

Table 5.14 Recall for 8-Class datasets

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Unigrams	0.22	0.24	0.24	0.14	0.19	0.24	0.21	0.19
Setup A	0.26	0.24	<b>0.22</b>	0.25	0.26	0.30	0.27	0.27
Setup B	0.30	0.26	0.27	0.24	0.26	0.29	0.32	0.30
Setup C	0.31	0.26	0.27	0.24	0.25	0.30	0.32	0.31
Setup D	0.31	0.26	0.27	0.24	0.24	0.29	0.33	0.32
Setup E	0.36	0.28	0.28	<b>0.34</b>	0.25	0.29	0.28	0.29
Setup F	0.32	0.28	0.27	0.26	0.28	0.31	0.31	0.29
Setup G	0.29	0.25	0.25	0.24	0.25	0.27	0.27	0.28
Setup H	0.26	0.26	0.26	0.24	0.22	0.27	0.25	0.25

Table 5.15 Recall for 3-Class datasets

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Unigrams	0.34	0.45	0.48	0.00	0.29	0.45	0.25	0.29
Setup A	0.48	0.34	0.48	0.00	0.37	0.56	0.37	0.22
Setup B	0.48	0.32	0.47	0.00	0.38	0.56	0.38	0.25
Setup C	0.47	0.32	0.47	0.00	0.38	0.57	0.39	0.24
Setup D	<b>0.24</b>	0.31	0.46	0.00	0.39	0.57	0.45	0.24
Setup E	0.48	0.31	0.46	0.00	0.39	0.57	0.39	0.26
Setup F	0.48	0.31	0.45	0.00	0.39	0.56	0.39	0.21
Setup G	0.47	0.35	0.47	0.00	0.37	0.57	0.38	0.22
Setup H	0.41	0.28	0.42	0.00	0.30	0.48	0.31	0.19

The 3-Class dataset with 2 emotions (i.e. Amused and Boredom) results of recall for the Setups are presented in Table 5.15. We can notice that the results vary, NB, LIN, ME and SMO showed an increase in recall in comparison with the unigrams alone, except for Setup D in the NB classifier, where there was a 10% decrease in recall (Bold in the Table). This could be due to the addition of the number of punctuation signs which was added in Setup D. In contrast, Yassine et al. [184] stated that number of punctuation signs are valuable in emotion detection. We noticed that MNB, CNB and RF showed a decrease in recall, except in Setup A from the CNB classifier. As for the RB classifier the recall remained 0%, which suggests that this classifier is not good at identifying emotion even with the extra features. Similar to the 8-Class dataset, we found the best experiment for the 3-Class dataset was Setup E, which consists of Positive and Negative Polarity, Number of Hashtags, Sarcasm, Number of Punctuation signs and Neutral Polarity. This is different from the all emotions Setup E by including sarcasm instead of number of emoticons. Similar to the 8-class dataset, Setup H which consists of the sarcasm feature led to the worst performance.

Table 5.16 Recall for Amused dataset

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Unigrams	0.25	0.40	0.57	0.00	0.29	0.28	0.12	0.11
Setup A	0.34	0.40	0.54	0.00	0.25	0.30	0.15	0.15
Setup B	0.34	<b>0.45</b>	0.57	0.00	0.32	0.23	0.16	0.14
Setup C	0.42	0.21	0.46	0.00	0.32	0.45	0.32	0.07
Setup D	0.44	0.20	0.42	0.00	0.34	0.46	0.33	0.09
Setup E	0.44	0.20	0.40	0.00	0.33	0.45	0.33	0.08
Setup F	0.34	0.37	0.47	0.00	0.31	0.37	0.18	0.16
Setup G	0.33	0.40	0.53	0.00	0.25	0.30	0.15	0.20
Setup H	0.33	0.39	0.50	0.00	0.31	0.35	0.18	0.19

The results of recall for the 2-Class Amused emotion dataset Setup results are presented in Table 5.16. The results varied, NB, LIN, ME, SMO and RF revealed an increase in the recall results, except from 6 cases which are Setups A and G the LIN classifier, Setup B in the ME classifier and Setups C, D and E in the RF classifier, where it led to a decrease in emotion identification. MNB and CNB classifiers showed a decrease in recall when adding features, except in Setup B for the MNB classifier where there was an increase in recall. Similarly to the 8-Class dataset, there was no increase in RB recall. We found that Setup B led to the best performance for the 2-Class Amused dataset across all the classifiers except for the ME classifier. We found the least useful Setup for this dataset was Setup F, which was all the features included.

As for the 2-Class Bored emotion dataset, the results of recall for the Setup results are presented in Table 5.17. The results varied, NB, LIN, ME, SMO and RF performances were increased by the Setups in comparison with the unigram feature alone. There was a decrease in MNB and CNB classifiers' recall when adding features, except in Setup F where in contrast of RF, there was an increase in recall. Similarly to the other datasets, RB showed no increase in recall. In opposition to the Amused dataset, we found the best results were with Setup F, except from the RF classifier, where it led to a decrease in emotion identification.

The 2-Class Excited emotion dataset recall results for the Setups are presented in Table 5.18. Classifiers NB and MNB alone, led to an increase in recall when adding features. In contrast classifiers CNB, LIN, ME, SMO and RF all revealed a decrease in recall, except for a few cases which are Setups C, D and E in CNB, ME and SMO. Similar to the rest of the datasets, RB showed no increase in recall. Additionally, similar to the 8-Class and 3-Class datasets, we found that Setup E led to the best performance, which consisted of Positive and Negative Polarity, Number of Hashtags, Time-End, Sarcasm and Neutral Polarity. In contrast to other datasets, this dataset consisted of the Time= E feature, which was ranked as

Table 5.17 Recall for Bored dataset

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Unigrams	0.44	0.53	0.63	0.00	0.38	0.41	0.29	0.41
Setup A	0.52	0.50	0.60	0.00	0.41	0.52	0.38	0.42
Setup B	0.54	0.50	0.59	0.00	0.46	0.55	0.41	0.47
Setup C	0.53	0.50	0.59	0.00	0.45	0.56	0.40	0.46
Setup D	0.55	0.50	0.57	0.00	0.46	0.56	0.40	0.43
Setup E	0.55	0.50	0.58	0.00	0.46	0.56	0.41	0.45
Setup F	0.55	0.68	0.72	0.00	0.44	0.65	0.42	0.28
Setup G	0.52	0.50	0.59	0.00	0.42	0.51	0.38	0.43
Setup H	0.46	0.48	0.54	0.00	0.39	0.46	0.33	0.42

Table 5.18 Recall for Excited dataset

Machine learning technique	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Unigrams	0.24	0.42	0.60	0.00	0.25	0.39	0.11	0.20
Setup A	0.35	0.42	0.57	0.00	0.22	0.29	0.10	0.18
Setup B	0.38	0.46	0.58	0.00	0.24	0.35	0.14	0.19
Setup C	0.38	0.78	0.80	0.00	0.17	0.56	0.19	0.05
Setup D	0.37	0.77	0.80	0.00	0.16	0.56	0.19	0.06
Setup E	0.38	0.78	0.80	0.00	0.16	0.56	0.18	0.08
Setup F	0.39	0.48	0.52	0.00	0.24	0.39	0.24	0.17
Setup G	0.38	0.42	0.56	0.00	0.22	0.29	0.10	0.17
Setup H	0.24	0.42	0.54	0.00	0.25	0.25	0.09	0.14

one of the top 6 features. Like the 8-Class and 3-Class datasets, Setup H led to the worst performance.

### Machine learning technique results

To find the best classifiers, we looked at the classifier leading to the highest results across each of the Setups for each dataset. As for the 8-Class dataset presented in Table 5.19, we found that the best classifier was SMO leading to the best results in Setups A-D. Similarly, Chaffar and Inkpen [26] research revealed that SMO led to a high performance. The second best classifier was NB leading to the best results in Setups E-G. This is similar to research by Tian et al. [170] and Danisman and Alpkocak [41], where NB led to a good performance. Danisman and Alpkocak [41] experimented with a multi-class model with 5 emotions and found that the Naive Bayes performed well, leading to an accuracy of 67%. We also found that for Setup H, ME led to the best performance (emotion detection). We found the highest recall was in Setup E, with the NB classifier. The highest AUC occurred in Setups F and G.

Table 5.19 Best classifier for the 8-Class dataset

Setup	Setup A	Setup B	Setup C	Setup D	Setup E	Setup F	Setup G	Setup H
Highest Classifier	SMO	SMO	SMO	SMO	NB	NB	NB	ME
Accuracy	0.27	0.32	0.32	0.33	0.36	0.32	0.29	0.27
Error rate	0.73	0.68	0.68	0.67	0.64	0.68	0.71	0.73
Precision	0.24	0.29	0.30	0.30	0.37	0.32	0.28	0.27
Recall	0.27	0.32	0.32	0.33	0.36	0.32	0.29	0.27
F-score	0.25	0.30	0.31	0.31	0.36	0.31	0.28	0.24
AUC	0.66	0.67	0.66	0.67	0.65	0.69	0.69	0.62

Table 5.20 Best classifier for 3-Class dataset

Setup	Setup A	Setup B	Setup C	Setup D	Setup E	Setup F	Setup G	Setup H
Highest Classifier	ME	ME	ME	ME	ME	ME	ME	ME
Accuracy	0.50	0.51	0.51	0.50	0.50	0.48	0.47	0.49
Error rate	0.50	0.49	0.49	0.50	0.50	0.52	0.53	0.51
Precision	0.35	0.34	0.35	0.35	0.35	0.35	0.35	0.33
Recall	0.56	0.56	0.57	0.57	0.57	0.56	0.57	0.48
F-score	0.43	0.42	0.43	0.43	0.43	0.43	0.43	0.39
AUC	0.64	0.64	0.64	0.64	0.64	0.62	0.64	0.62

The 3-Class dataset results illustrated in Table 5.20 showed that best classifier was ME which led to the best results in all the Setups. Our results are higher than Wang et al. [179] study that used a multi-class model with the ME classifier, achieving an accuracy of 35%. On the other hand, their research explored detecting emotions from Chinese blogs. In addition, they only presented the accuracy and did not include the recall which is needed to assess the detection of the emotion class.

The best classifier for the 2-Class Amused dataset presented in Table 5.21 was CNB, which led to the best performance in all Setups except Setup D and E, where ME led to the highest performance. Additionally, we found that CNB led to the best model in terms of recall (i.e. 57%) with Setup B. Our results could suggest that CNB was the best classifier due to the unbalanced classes in the 2-Class dataset (i.e. most of the data in the 'other' class), and based on our previous findings CNB addresses the uneven data classes' problem. However, the models which identified the emotion best have a low accuracy. This suggests that the model is incorrectly classifying the other class as emotion.

As for the 2-Class Bored and Excited datasets presented in Table 5.22 and Table 5.23, we found that CNB performed best for all the Setups. We found the highest accuracy for



Table 5.21 Best classifier for Amused dataset

Setup	Setup A	Setup B	Setup C	Setup D	Setup E	Setup F	Setup G	Setup H
Highest Classifier	CNB	CNB	CNB	ME	ME	CNB	CNB	CNB
Accuracy	0.69	0.70	0.75	0.74	0.74	0.69	0.68	0.68
Error rate	0.31	0.30	0.25	0.26	0.26	0.31	0.32	0.32
Precision	0.25	0.27	0.31	0.29	0.29	0.24	0.25	0.25
Recall	0.54	0.57	0.46	0.46	0.45	0.47	0.53	0.50
F-score	0.34	0.37	0.37	0.35	0.35	0.32	0.34	0.33
AUC	0.62	0.62	0.64	0.71	0.71	0.60	0.62	0.58

Table 5.22 Best classifier for Bored dataset

Setup	Setup A	Setup B	Setup C	Setup D	Setup E	Setup F	Setup G	Setup H
Highest Classifier	CNB	CNB	CNB	CNB	CNB	CNB	CNB	CNB
Accuracy	0.66	0.67	0.67	0.67	0.67	0.50	0.67	0.65
Error rate	0.34	0.33	0.33	0.33	0.33	0.50	0.33	0.35
Precision	0.37	0.38	0.38	0.38	0.38	0.39	0.38	0.35
Recall	0.60	0.59	0.59	0.57	0.58	0.72	0.59	0.54
F-score	0.46	0.46	0.46	0.46	0.46	0.46	0.46	0.43
AUC	0.64	0.64	0.64	0.64	0.64	0.57	0.64	0.61

emotion detection was 60% for the Bored model with Setup A. As for the Excited model, the highest recall was 80% found in Setups C, D and E. Oppositely in Gilbert and Karahalios [62] research about estimating emotions (anxiety, worry and fear) from 20 million posts weblogs found that Complement Naive Bayes algorithm results were rather low, giving an accuracy of only 32%.

### 5.3.4 Discussion

This research looked into identifying models for emotion detection, in addition to exploring the role of features. The results of our experiments indicate that adding extra features improves the emotion detection compared with using the unigrams alone.

Some of the features required extra input from the student and might not be realistic in real life settings, for instance the time of lecture features. However, if the information is supplied and available, the emotion identification for the models are significantly increased.

The number of hashtags were found useful in the feature usefulness experiment. This may indicate that hashtags in students' feedback are used for the purpose of expressing emotion. In contrast, we found that using the number of emoticons is not useful, despite that

Table 5.23 Best classifier for Excited dataset

Setup	Setup A	Setup B	Setup C	Setup D	Setup E	Setup F	Setup G	Setup H
Highest Classifier	CNB	CNB	CNB	CNB	CNB	CNB	CNB	CNB
Accuracy	0.71	0.72	0.55	0.55	0.55	0.73	0.72	0.70
Error rate	0.29	0.28	0.45	0.45	0.45	0.27	0.28	0.30
Precision	0.24	0.25	0.19	0.20	0.20	0.24	0.24	0.23
Recall	0.57	0.58	0.80	0.80	0.80	0.52	0.56	0.54
F-score	0.34	0.35	0.31	0.31	0.31	0.33	0.34	0.32
AUC	0.65	0.66	0.66	0.66	0.66	0.66	0.65	0.63

previous studies have generally found it useful for emotion detection. As already mentioned in chapter 4, this could be due to the small number of emoticons appearance in the dataset (i.e. only 63 instances out of our whole dataset).

We found that the Positive and Negative lexicon-based features are more useful than the Neutral lexicon-based feature. This was expected due to the Neutral-based features size being only 16% of the data. As already mentioned in chapter 4, the neutral class is also less common in previous research.

Munezero et al. [118] aimed to detect emotions from students' diaries, however their research was using a lexicon-based approach only. They did not report of any evaluating metrics for their models, therefore it is difficult to compare our results with theirs.

Another research used machine learning to predict emotions in an educational context is Tian et al. [170]. They used POS-tagging as features. From the emotions that we identified as relevant for learning from previous literature, they only looked at anxiety, for which they obtained a precision value of 0.6 using a LogitBoot classifier. This research, however, was conducted on Chinese text, which has different characteristics and structures compared with English text, and was also based on text from online chats and discussion groups.

Other research on emotions using machine learning looked at predicting the presence of emotion versus the lack of emotion [4]; while the maximum accuracy obtained was around 74%, the authors did not discuss the performance in terms of identifying the presence of emotion.

In this experiment, we use different features that have not been used in the educational domain before. Most of the research in emotion prediction has focused on common features such as n-grams and POS-tagging. However, our research suggests the possibility of increasing emotion prediction accuracy by adding extra features.

## 5.4 Optimal Model for Detecting Students' Emotion

Results from experiments 1 and 2 suggested that the single emotion detection models lead to the highest performance. Consequently, they were chosen for application. Moreover, our results indicated that three emotions (i.e. Amused, Bored and Excited) could be detected easier than others. These could be chosen as the only reliable models for our application. As for the models components, our results indicate that preprocessing decreases the performance of the models with our kind of data, therefore, we chose to eliminate it and only include a low level preprocessing with tokenization and removal of numbers and punctuation and stop words. We chose unigrams as a feature and CNB as a classifier due to their acceptable performance.

## 5.5 General Discussion

Emotion detection is important in education as they can have effect on learning. After reviewing different methods to detect emotion, we found that detecting emotions from text is a less costly method as it requires no extra equipment. Consequently, we proposed detecting emotions using sentiment analysis from students' textual feedback. The literature showed there are not many studies investigating the detection of emotion via text, hence the purpose of this study.

In experiment 1, we explored models detecting multiple emotions and single emotions. By observing other studies in emotion detection we found the majority of them used accuracy to evaluate their models. This can be misleading due to the existence of the 'other' class in the evaluation. On the other hand, recall indicates the percentage of correctly identified instances for a class of interest and it can be used to assess the ability of the classifiers to predict emotions. In addition, precision can refer to where the identification problems occur.

We found that the multiple emotion models were high in accuracy and low in recall. This implies that the models' accuracies are affected by the 'other' class, as the recall is low or even 0% in some cases. In addition, precision is also low (with a few exceptions). This indicates that the high accuracy is due to the correct identification of the 'other' class rather than the correct identification of emotion(s) and that the models are not useful for emotion detection. In contrast, we found that models detecting a single emotion were low in accuracy and high in recall. We found that these models are much more suitable for detecting emotion. We found three emotions could be detected more reliably than others (i.e. Amused, Bored and Excited). These were chosen for our application (i.e. the system described in chapter 8).

As for preprocessing the data, we found from experiment 1 that preprocessing led to a decreased performance. The negative influence of preprocessing on the performance of the models indicates that information typically discarded for polarity models have value for the identification of specific emotions.

By observing results from experiment 1, we found unigrams and trigrams led to the best performance for the 2-Class models, while unigrams combined with bigrams and trigrams led to the best performance for the multi-class models. In experiment 2, we reviewed other features that could be used in emotion detection which are domain-related features (i.e. lexicon-derived and time of the lecture), sarcasm, Twitter-related features (i.e. number of hashtags and emoticons), punctuation-related features (i.e. number of punctuations signs and question and exclamation marks) in combination with unigrams. We found that these features improved the results of the models in comparison with the unigrams alone. However to apply some of these features (e.g. sarcasm) in real-life settings is unrealistic. For our application, unigrams were chosen as the best feature due to them leading to a high performance in single models (i.e. Amused, Bored and Excited). Also, it is best to choose unigrams as it is preferable to use the least number of features possible to obtain the results due to the lower complexity of the model which has an implication on the processing time.

The machine learning techniques explored and commonly used for emotion detection are the same ones found in polarity detection. In experiment 1 and 2, we found that CNB was the best classifier. This could be due to its nature in handling unbalanced classes, which is the case in our dataset as it consisted of a small distribution of data for each emotion. Across both experiments, MNB, ME and SMO classifiers performed well for emotion detection.

In general, our research suggests that emotions could be detected from text in educational settings. The results of the emotion detection are promising in comparison with other studies in emotion detection.

In comparison with other studies that detected emotion using other methods such as facial expressions, speech and body gestures, we found that our results were considerably lower than Busso et al. [20] and Terzis et al. [164]. Terzis et al. [164] was only successful at detecting the anger emotion. Our research was on the same level as Arroyo et al. [8] and Dellaert et al. [45]. To conclude, using other methods may be more accurate, but using text based method may be more feasible in classroom settings as it does not require extra expensive equipment.

This study has several limitations. Firstly, the dataset was small, and the data for each emotion class was imbalanced. As previously mentioned in chapter 4 previous literature has revealed the difficulty of collecting educational data. In addition, due to the small datasets, some of the features such as emoticons and number of question and exclamation marks were

limited, which may have the reason to their small effect in performance. We also found that although some of the features were beneficial to the models' performance (i.e. sarcasm), they would require extra labelling from the students.

## 5.6 Summary and Contribution of the Chapter

In this chapter, emotion detection was explored. We reviewed the the literature related to emotion detection and highlighted its importance in education. We presented the benefits of detecting emotion from text over standard methods such as facial and speech emotion detection from camera and sensors.

This chapter presents two experiments: n-gram feature exploration and other feature exploration. By observing the results, we found that preprocessing data leads to a worse performance due to it removing valuable elements for the detection.

Our results indicate that models which detect a single emotion have a better performance than multiple emotion models. Additionally, three single emotion models (i.e. Amused, Bored and Excited) were more reliable at detecting emotion.

Experiment 1 results suggest that unigrams and trigrams led to the best performance for the 2-Class models, while unigrams combined with bigrams and trigrams led to the best performance for the multi-class models. As for experiment 2, using 15 additional features led to the highest performance in the majority of models.

The best classifier across both of the experiments is CNB which could be due to the imbalanced nature of the data. MNB, ME and SMO, performed well, however, CNB addresses the imbalanced datasets problem, therefore was selected for the application.

We presented the optimal models to detect emotion, which were models detecting single emotions. The models consisted of low level preprocessing, unigrams as features and CNB as the classifier.

In comparison to other studies in emotion detection, our study reveals promising results that are as effective as emotion detection using other techniques. In contrast to other studies in emotion detection, we use various machine learning techniques and features that have not been explored before. This research directs future researchers into looking at the role of features more, as the features can improve performance of emotion detection.

The main contribution of this chapter was to explore machine learning algorithms to detect students' classroom emotions. This contribution is spilt into several contributions:

- The most common learning emotions were extracted from previous literature on emotions in education. These were 19 emotions of which eight of them were found

to be the most frequent in students feedback (i.e. Bored, Amused, Frustrated, Excited, Enthused, Anxiety, Confused and Engaged).

- Our study is the first to explore preprocessing in emotion detection. Experiments with different preprocessing levels indicate that a lower level of preprocessing is more useful for emotion detection from text.
- This study explored several features for emotion detection in students' feedback. This included exploration of less common features and their combinations that have not been used before in emotion detection. Using other features in combination with unigrams increased the performance of the models.
- Several machine learning techniques (i.e. CNB, SMO, RF) that are less common in emotion detection were experimented with. We found that CNB performed well with models which detect a single emotion which could be due to the unbalanced classes.

# Chapter 6

## Detecting sarcasm from students' feedback

In this chapter we explore methods to detect students' sarcasm in their feedback. We review previous research related to sarcasm detection from text and then present our experiments and results.

### 6.1 Sarcasm Detection

Sarcasm is a form of irony conveying criticism in an indirect way [150]. It may also be used as a way of covering up embarrassment in a way that is dramatic or humorous [151]. Unlike lying, sarcasm is used to highlight reality rather than hide it. It is perceived as being more polite and less aggressive than direct confrontation [151].

Sarcasm has been extensively studied in linguistics, psychology and cognitive sciences, e.g. [150, 151, 176, 178]. From neuroscience studies in particular, it has been shown that to recognise sarcasm, contextual cues are essential. However, when contextual cues are missing, sarcasm can also be detected from voice prosody and facial expression [151]. Given the lack of such cues in text, sarcasm detection can only rely on the contextual cues. In addition, the subtle language used sometimes to convey sarcasm makes it difficult to detect from text only [130].

In the context of education, sarcasm has been studied from the perspective of the teacher/lecturer. More specifically, there have been studies investigating the use of sarcasm in communication or providing feedback to students [21, 143, 171]. These studies suggest that sarcasm can be negative when used to belittle students and positive when used

with humour and not targeted at particular students. To the best of our knowledge, there is no research investigating sarcasm in students' textual feedback.

As pointed out previously, detection of sarcasm from text is a challenging task and there are only a few studies exploring this problem [42, 67, 80, 99, 140, 175], which will be reviewed in the following paragraphs. None of these studies, however, explore sarcasm in the educational context, which is different from typical social media interaction.

Detection of sarcasm has received increased attention from the research community, with several works reported for the detection of sarcasm from text only [42, 67, 80, 99, 140, 175].

Davidov et al. [42] investigated the prediction of sarcasm from general Twitter data, as well as Amazon product reviews. They investigated the use of several features, including punctuation signs and '#sarcasm' hashtag identification and the use of capital letters. They collected 5.9 million tweets collected from Twitter, and 66000 Amazon product reviews. They employed the seed method to automatically extend the small number of manually labelled instances; 80 seeds (i.e. 80 sentences labelled as sarcastic) were used for both Twitter and Amazon datasets. For the Amazon data, 471 sarcastic and 5020 non-sarcastic sentences were obtained as the training set. For the Twitter dataset, the authors did not report the number of sarcastic and non-sarcastic tweets in the training set. They performed two experiments i.e. one on Twitter and one on Amazon data; for each experiment they used a test set of 90 sarcastic sentences and 90 non-sarcastic sentences (sampled from the whole corpus). An F-score of 83% was achieved using a k-nearest neighbour classifier for the Amazon dataset and 54% for the Twitter dataset.

González-Ibáñez et al. [67] distinguished sarcastic tweets from positive and negative ones using a lexicon-based method. Their dataset consists of 900 tweets in each of the three categories, sarcastic, positive and negative. They explored the contribution of different lexicons on the performance of classifiers in detecting sarcasm; the results for accuracy varied between 55% and 75%. In addition, they compared the performance of their models and lexicons against human labelling and found that it led to a similar performance.

Justo et al. [80] focused on the detection of sarcasm and nastiness from online forums conversation data. This study used the Internet Argument Corpus (IAC) as a dataset which is a collection of 390704 posts in 11800 discussions extracted from the online debate site 4forums.com. They used a Multinomial Naive Bayes classifier and investigated a variety of features; a 70% F-score was achieved for sarcasm and 79% for nastiness.

Liebrecht et al. [99] investigated the detection of sarcasm from Dutch tweets. The dataset of this study consisted of around 78083 sarcastic tweets and over 3 million non-sarcastic ones. They used Winnow classification as a machine learning technique. Highly imbalanced classes were used and consequently they found that sarcasm could be detected with true



positives rates of 75% for balanced data and of 56% for imbalanced data. Their results showed that the strongest linguistic indicators of sarcasm were broad synonyms of the word sarcasm, such as irony, cynicism and humour.

Riloff et al. [140] detected sarcasm by looking at the contrast between a positive/negative sentiment and a situation, such as positive sentiments expressed in relation to a negative situation. The authors collected 1600 tweets with a sarcasm hashtag (#sarcasm or #sarcastic), and 1600 tweets without these sarcasm hashtags from Twitter's random streaming API. They use part-of-speech (POS) tagging and n-grams as features and Support Vector Machines as a classifier, which led to 63% precision and 39% recall.

Tsur et al. [175] investigated sarcasm detection from product reviews. Semi-supervised learning was used to expand an initial set of labelled data. The data collected for this study was 66000 reviews for 120 products extracted from Amazon.com. Only part of the data was selected in the experiment due to the labelling process being manual. They used 80 sarcastic sentences and 505 non-sarcastic sentences (from 80 non-sarcastic reviews) as seeds, which led to a training set of 471 sarcastic instances and 5020 non-sarcastic instances. For the testing set, they used 90 sarcastic and 90 non-sarcastic sentences (this is the same as the Amazon data that was used by Davidov et al. [42]). They performed two experiments according to the level of sarcasm in the sentence. This study focused on pattern-based features; achieving a F-score of 83% with k-nearest neighbour. The authors also experimented with punctuation-based features and found that they could contribute to a better classification performance.

In the research reported above, all used 2-class models, i.e. sarcastic and non-sarcastic, with the exception of González-Ibáñez et al. [67] who experimented with a 3-class models, i.e. positive, negative and sarcasm. For our experiment, we use a 2-class models.

Unlike previous research, our experiments focus on data from an educational context and more specifically, from students' feedback. We also experiment with several classifiers to investigate the reliability of prediction. In the following paragraphs, we review in more detail the previous work in relation to prediction models in sarcasm detection.

Most of the research about sarcasm in sentiment analysis has been on tweets. Davidov et al. [42] replaced Twitter special characters with words; for instance each hashtag was replaced with the word 'HASHTAG'. Tsur et al. [175], on the other hand, removes Twitter special characters. Additionally, both of these studies remove the HTML links [42, 175]. Emoticons were used by González-Ibáñez et al. [67] in their study exploring polarity and sarcasm from tweets. Emoticons and hashtags have been also used as features in detecting sarcasm [42, 67, 175]; consequently in this study, they are experimented with as well.

In sarcasm, n-grams are commonly used as features [67, 99], although other features such as part-of-speech (POS) tags have been used [140].

The use of lexicon-based features is also popular for detecting sarcasm and has been used in various previous research, e.g. [67, 140]. Other studies looked into other features such as punctuation-based [42, 175], emoticons [67], quotes [42, 175] and the number of capitalised/all capital words in the sentence [42, 175].

A variety of *machine learning techniques* have been used for sentiment analysis in general and sarcasm detection in particular. Some of the common classifiers that have previously been used in sarcasm detection are Multinomial Naive Bayes (MNB) [80] and Support Vector Machines (SVM) [42, 140]. Other techniques which have previously been used in prediction models and could be promising for sarcasm detection are: Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME), Sequential Minimal Optimization (SMO) and Random Forest (RF).

## 6.2 Data Corpus

The data used in this experiment is described in chapter 4, in section 4.3.1. To identify and label sarcastic tweets two techniques were used: (a) the presence of the word or hashtag 'sarcasm' (or other variants such as 'sarcastic') in the tweet and (b) find if the tweets were labelled an emotion that was opposite of the tweet meaning by a human annotator, for example, a tweet about the lecture being boring and the label set as amused. Using these two methods, 194 sarcastic tweets were found; some examples of sarcastic tweets from both categories are presented in Table 6.1. For the purpose of classification, all the other tweets were labelled as non-sarcastic.

## 6.3 Predicting sarcasm from students' feedback

Similarly to the previous experiments, different levels of preprocessing were investigated. We tested three preprocessing levels using different techniques which were chosen due to their popularity in previous studies and used previously in chapter 4 in experiment 4.3.2:

1. Preprocessing P1: This is to test how well the model works without any preprocessing apart from converting the letters into lowercase. We included this to explore the role of different degrees of preprocessing, in which this one (P1) is the baseline;
2. Preprocessing P2: This is to remove numbers, punctuation, spaces and blanks and special characters. In most cases these do not hold value or sentiment, therefore they are noise to the data; and

Table 6.1 Examples of Sarcastic Tweets

Sarcastic tweets from students' feedback	Justification
What if i could not sleep and tweet for the first time? #english #lecture #sarcasm #goodnight	#sarcasm in the sentence
Nothing great than doing some vectors on sunday morning #vectors #lecture #sarcasm #just-universitythings	#sarcasm in the sentence
in need of some inspiration #literature #lecture #boring #feelingdown	labelled 'Amused'
the only time i didn't want the lecture to end #english101 #coollecturer #hadoop	labelled 'Bored'
i would love to have some coffee to end this misery! #english #lecture #boring	labelled 'Engagement'
finally got my ipod charged, time for some music #history #lecture #boringlecture	labelled 'Enthusiasm'
can't wait for calculus to be over #sad #bored	labelled 'Enthusiasm'
dr. jamal needs to seriously needs to speak up #chemistrylecture #frustrating #canthear	labelled 'Excitement'
x=whatagain? #lost #confused	labelled 'Excitement'

3. Preprocessing P3: In addition to the preprocessing in level P2, this includes the removal of hashtags, emoticons and twitter special characters.

Like most researchers, we focus on n-grams and in particular on unigrams. Unigrams have been used to detect sarcasm [67, 99]. We also explored other features in combination with unigrams (U):

1. emotion label (E) (from students- the 19 emotions described in chapter 5, section 5.1.4)
2. polarity label (P) (from students- positive, negative and neutral)
3. number of all punctuation characters (PC) (e.g. commas, semicolons, brackets, apostrophes, etc.)
4. number of question and exclamation marks (QE)
5. number of emoticons (NE)
6. number of hashtags (H)
7. time of the tweet during the lecture (T), (i.e. beginning, middle, end)

All combinations of these additional features with unigrams were experimented with.

As for the machine learning techniques used, we chose popular classifiers that have been used before in sarcasm detection: Multinomial Naive Bayes (MNB) and Support Vector Machines (SVM). We also use Naive Bayes (NB), Complement Naive Bayes (CNB), Maximum Entropy (ME), Sequential Minimal Optimization (SMO) and Random Forest (RF), due to their popularity in predicting models in general and due to their promising performance in your previous experiments (described in chapter 4 and 5).

## 6.4 Results and Discussion

All the models were tested using 10-fold cross-validation; accuracy, precision, recall, F-score and error rate are used to evaluate the models. Precision, recall, F-score and Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) for the sarcasm class are also reported, to assess the performance of the model for this class of interest. AUC of ROC illustrates the relative trade-off between true positives and false positives.

In relation to *preprocessing* the text, we first experimented with unigrams as features and found that most of the models showed higher performance in terms of sarcasm detection when using the lowest level of preprocessing (P1). The reason that most of the classifiers perform best at this level could be due to hashtags, numbers and punctuation holding value to the prediction of sarcasm. In previous work, punctuation signs identification was used by Tsur et al. [175] and Davidov et al. [42]. Consequently, for the unigrams combinations with the other features, we only experimented with level P1 of preprocessing.

From the *features* experimented with, we found that adding the emotion and polarity labels to the unigrams led to an increase of 1 to 6% in sarcasm detection, i.e. recall, in five out of the eight classifiers (i.e. NB, MNB, CNB, ME and SMO). We found that overall, the punctuation features (PC and QE), the emoticons count (NE) and the time of the lecture (T) made little difference to the performance and in some cases led to a small decrease of 1% to 2%. Using all the features led to an increase for the recall of the sarcasm class for the NB, LIN and SMO classifiers by 7%, 4% and 9%, respectively.

In terms of *machine learning techniques*, the best results for sarcasm detection for each machine learning technique are given in Table 6.2. We found on our data, the best performing classifiers were Complement Naive Bayes, due to its high recall rate. Multinomial Naive Bayes and Naive Bayes were considerably lower than CNB. Naive Bayes has not been explored before to detect sarcasm in sentiment analysis, however, previous research shows that it performs well in sentiment analysis [131, 136]. Multinomial Naive Bayes has been used in sarcasm detection by Justo et al. [80], obtaining a recall of 77%. We notice

Table 6.2 Best models for each machine learning technique in sarcasm prediction

	NB	MNB	CNB	RB	LIN	ME	SMO	RF
Features	All	U+E+P	E+P	U	All	E+P	All	U+H
Accuracy	0.82	0.78	0.72	0.87	0.80	0.85	0.86	0.84
Precision	0.84	0.84	0.84	0.76	0.82	0.84	0.84	0.81
Recall	0.82	0.78	0.72	0.87	0.80	0.85	0.86	0.84
F-score	0.83	0.80	0.76	0.81	0.81	0.85	0.84	0.82
Error rate	0.18	0.22	0.28	0.13	0.20	0.15	0.14	0.16
Precision (Sarcasm)	0.34	0.30	0.26	0.00	0.28	0.42	0.43	0.30
Recall (Sarcasm)	0.47	0.51	<b>0.63</b>	0.00	0.35	0.33	0.25	0.22
F-score (Sarcasm)	0.39	0.37	0.36	0.00	0.31	0.36	0.31	0.25
AUC	0.76	0.71	0.68	0.50	0.61	0.75	0.60	0.68

Table 6.3 Results from previous research

	[42] A	[42] B	[67]	[80]	[99] A	[99] B	[140]	[175]
Precision	0.91	0.72	-	65.30	-	-	0.63	0.91
Recall	0.76	0.44	-	77.40	0.75	0.56	0.42	0.76
F-score	0.83	0.54	-	70.70	-	-	0.51	0.83
Accuracy	0.95	0.90	0.71	67.90	-	-	-	0.95

that interestingly, CNB performs the lowest in terms of accuracy, despite previous research showing that it performs well with uneven classes [65], which is applicable to our dataset as the sarcasm class represents only 13% of the data; however, CNB led to the highest sarcasm recall of 63%.

We found that the lowest performing classifiers to detect sarcasm are RB, SMO and RF. Despite of the high accuracies in these classifiers, the recall and other measures for the sarcasm class were low, indicating that the good performance is due to the recognition of the majority class, i.e. non-sarcasm. Similarly, research by González-Ibáñez et al. [67] to identify sarcasm from tweets showed that SMO led to a high accuracy of 71% using unigrams as features. They did not, however, report the measures for the sarcasm class; consequently, we do not know how much of the accuracy performance is due to the recognition of sarcasm.

To compare our results with previous research, we display these results in Table 6.3. For two of the previous works, we report two sets of results which are relevant: Davidov et al.[42] A and B represent their results on Amazon and Twitter data, respectively; Liebrecht et al.[99] A and B represent their results on balanced and unbalanced classes, respectively.

One aspect to consider when looking at these results is the data that was used for analysis. In fact, the best models were actually developed for data other than Twitter. Both Davidov et al. [42] (A) and Tsur et al. [175] report results on Amazon reviews - in fact the results

reported by the two papers are the same. Davidov et al. [42] also report results on Twitter data, i.e. [42] B, which are much lower. Justo et al. [80] used data from online forum conversations on social and political topics. This distinction between Twitter data and other types of data is important because Twitter data has characteristics that are not present in other types of data; in addition, the research so far indicates that detecting sarcasm from Twitter data is more difficult than other online data – this is also supported by the summary results of Table 6.3. This may be due to the shortness of tweets, which does not allow for rich contextual information, which is essential in the detection of sarcasm [151].

All of the previous research looking at detection of sarcasm from text, with the exception of Liebrecht et al. [99], report the overall results for the classifiers, without discussing the performance of those classifiers for the sarcasm class, which is the one of interest. Consequently, we do not know if the overall performance is due to a classifier being able to detect the non-sarcasm class, which in many cases is the majority class. In other words, we do not know if those classifiers are good at detecting sarcasm rather than non-sarcasm. If we were to judge our results based on the overall performance, five out of the eight the classifiers we investigated had all the metrics reported in Table 6.3 (i.e. accuracy, precision, recall and F-score) above 0.80. Compared with previous research, this would indicate a very good performance, even when comparing with non-Twitter data.

The results for the sarcasm class, however, are the only ones that can clarify a classifier's ability to detect sarcasm. From previous research, we can only compare with the results on Liebrecht et al. [99], which in Table 6.3 are given for the sarcasm class. When using balanced data, i.e. [99] A, they achieved a very good recall of 0.75; however, on unbalanced data, i.e. [99] A, the recall is 0.56. They did not report the other indicators in the table. As our data is unbalanced, a fair comparison should be made only with their results on unbalanced data. One of our classifiers, i.e. Complement Naive Bayes, achieved a better recall of 0.63.

Our results as well as the findings of previous research confirm that detection of sarcasm from text is a challenging task that needs further investigation. Although the results we reported in this study are relatively low for what is considered a good performance for a classifier, when looking at the previous results on sarcasm detection, we can see that our results are comparable and on some metrics even better than previous research.

From our data, we found that sarcasm is used in text in educational context. Our results could direct other researchers in the sentiment analysis community to explore labels in supervised datasets in other domains as well.

One limitation to this chapter is that our data is small and only 196 sentences were sarcastic. Most of the studies have explored sarcasm in tweets with larger datasets. However,

the purpose of this study was focused on students' feedback and as mentioned earlier in chapters 4 and 5, the data is difficult to obtain.

## 6.5 Summary and Contribution of the Chapter

In this chapter, we explored the detection of sarcasm in students' feedback. We reviewed previous literature and found that sarcasm is a problem in sentiment analysis and particularly in the educational domain due to students using it in their feedback.

Two methods were used to label students' sarcasm: mismatched labels and detecting 'sarcasm' and 'sarcastic' words and hashtags. We used two classes: sarcastic and non-sarcastic.

Three preprocessing models were explored. We found that the lowest preprocessing level (P1) resulted in the highest performance. This indicates that hashtags, numbers and punctuation are valuable in sarcasm detection.

We explored unigrams alone and with 7 other features. Our results showed that adding emotion and polarity labels leads to an increase of 1 to 6% in sarcasm detection (i.e. recall). Using a combination of all the features led to an increase in sarcasm detection in the NB, LIN and SMO classifiers by 7%, 4% and 9% respectively.

As for the machine learning results, we found the best classifier is CNB, which led to the best recall of 63%. We found that the lowest performing classifiers were RB, SMO and RF, due to their low recall which reflects sarcasm detection.

The contribution of this chapter was to explore sarcasm detection in students' feedback. The contribution includes an exploration of different features that have not been used before such as emotion and polarity labels and time of the tweet. We found by other features such as emotion and polarity labels led to an increase of 1 to 6% in sarcasm detection (i.e. recall). We also investigated the performance of different classifiers that have not been used in sarcasm detection before (i.e. NB, CNB, SMO, ME and RF). We found that the best classifier was CNB and the lowest performing classifiers were RB, SMO and RF.





# Chapter 7

## Visualisation

This chapter presents the research conducted regarding the lecturers' preferences about how the results from the sentiment analysis models should be visualised. Section 7.1 presents previous research about visualisation in general and particularly in sentiment analysis and emotion detection. Section 7.2 describes the experiment including information about the visualisations, participants and the procedure of the experiment. Section 7.3 explains the results followed by discussion in section 7.4. A summary, the contribution and the conclusion of this chapter are presented in section 7.5.

### 7.1 Previous research

Visualisation is used to represent data, making it easier for the user to understand and interpret. There are many types of visualisation graphs according to the type of data. From the most well-known visualisations are bar, pie, line, scatter and area charts which are available in programs such as Microsoft Office and Weka. These basic charts attracted the attention of researchers and have been studied from different perspectives [37, 106, 145, 147, 148, 172]. They are usually used to represent numerical data. Scatter plots are used to explore the possible relationship between two variables relating to the same event. They are normally used to present data in Weka, the data mining tool [73]. There are also visualisations to represent words; for instance wordle is a tag cloud of words of different sizes reflecting their importance and frequency [88]. In sentiment analysis, words could highlight the main reasons of the polarity/emotion results, therefore it is important that the chosen visualisations include frequent words.

Visualisations have different advantages and disadvantages depending on the application and audience. One of the problems in complex visualisations is interpreting them incorrectly. For example, the study by Elting et al. [52] showed that physicians could make wrong

decisions depending on the visualisation which could have major consequences. Additionally, presenting the data wrongly can cause confusion [52]. The most common visualisations advantages and disadvantages from the literature [37, 145, 147, 148, 172] are highlighted in Table 7.1.

Table 7.1 Visualisation advantages and disadvantages

Graph	Advantages	Disadvantages
Bar Chart	Shows the data distribution. It is a simple chart and easy to interpret and it can provide frequencies and it easy to detect the minimum and maximum values/range/outliers.	It is not suitable for changes over time or for making predictions.
Box and Whisker Chart	Summarises large amounts of data, it shows outliers and the distribution of data	Individual data instances and relationships between variables are not shown. Additionally, it cannot compare categorical data.
Histogram Chart	It is visually strong and categories/numbers can be compared easily	Does not read individual values because data is grouped into categories. It is difficult to compare between two data sets. Additionally, it is only suitable with continuous data.
Line Chart	Can compare multiple continuous data sets easily and data that changes over time.	It can only be used with continuous data and is not useful for comparing categorical data.
Pie Chart	Displays percentages and performs well with large numbers.	It cannot be used with a large number of categories.
Scatter Chart	It shows the exact data values and sample size. It also shows the minimum/maximum and outliers.	It is not suitable for visualising large data sets. In addition, the data on both axes should be continuous.

Visualisation has been used previously to present sentiment analysis results. For example, Marcus [106] used pie charts to present sentiment analysis results from Twitter data.

Neri et al. [119] used bubble maps to present sentiment analysis clustering in social media (See Figure 7.1). They also used pie charts to present polarity and line charts to illustrate polarity over periods of time.

Visualisation has also been used to present emotional data. One study highlights that emoticons can be to used describe emotional data [25]. The study was done by Callele et al. [25] who used emoticons to identify players' emotional state in a virtual world (i.e. gaming).

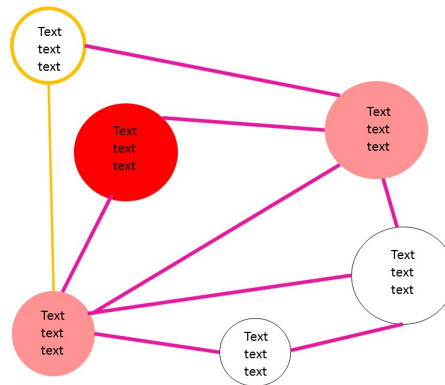


Fig. 7.1 Neri et al. [119] bubble plot

Gregory et al. [69] created a new visualisation for automatic sentiment analysis. This visualisation was originally adapted from the box plot which they named rose plot due to the shape being similar to the rose petals (See Figure 7.2). Their visualisation included detection of four pairs of affect categories:

1. Positive (n=2236)-Negative (n=2708)
2. Virtue (n=638)-Vice (n=649)
3. Pleasure (n=151)-Pain (n=220)
4. Power Cooperative (n=103)-Power Conflict (n=194)

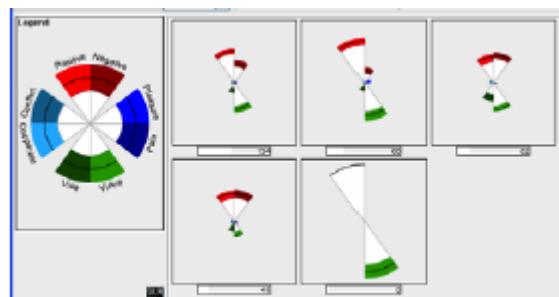


Fig. 7.2 Gregory et al. [69] rose plot

They used the size of the petals in the rose plot to illustrate the larger class category. One advantage of their research is providing the user with polarity and emotion integrated in a single visualisation.

As sentiment analysis is related to word analysis, we also explore methods to visualise text. As mentioned previously, text can be presented in a word cloud, such as Wordle. Wordle

scales words according to their frequency and usually in many cases, the more frequently they appear in the text, the more important the words are [88]. Mcnaught and Lam [110] highlighted that word clouds could be a useful research tool to aid educational research such as providing extra support (i.e. visualisation) for learning analytic tools. They also highlighted that word clouds allowed viewers to quickly visualise general patterns in text.

Knight [87] showed that the domain has an influence on visualisation. Acquiring knowledge about the domain can save time, due to the similarities in visualisation tasks and type of data [87]. In addition, the visualisations are usually specific to the type of data [87].

Visualisations have been used in many domains including education. Geryk [61] created an analytical tool to increase education efficiency and quality. They explored different visualisations which enable better exploratory of interactive analysis. They visualised student data, including their grades and the unit credits to find students that drop-out or change course. They used bubble charts which is to represent data as bubbles, where the size of the bubbles increase according to the data.

Romero et al. [142] created a data mining and visualisation tool for the discovery of student trails in web-based educational systems for lecturers to use. The visualisation for this tool was a graph with nodes and edges. Each node represents a web page and the directed edges (i.e. arrows) indicate how the students have moved between the web pages.

Leong et al [94] used students' feedback via SMS messages to evaluate teaching. The messages were classified into positive and negative classes, which were visualised to the lecturer. They created a complex map of positive and negative terms with links between them; the lecturer would however, have to estimate the proportions of positive and negative concepts, which would be time consuming.

The visualisation in our case is specialised in representing students' polarity and emotion from students' feedback to the lecturer. To the best of our knowledge, the extensive review of the literature did not reveal any suitable visualisation of students' feedback in the context of a classroom environment that presents polarity and emotion to the lecturer.

## 7.2 Experiment

The purpose of this experiment is to investigate the best graph(s) to visualise the results of sentiment analysis (i.e. polarity and emotion detection) on students' real-time feedback to the lecturer. The visualisation is an important component that should facilitate the lecturers understanding of the class opinions through simple displays that highlight meaningful information. We chose a participatory design for this experiment to allow lecturers to participate in providing feedback on suggested visualisation designs, as well as allow them

to suggest other designs. The following subsections include the participants, the materials which describe the visualisation designs and the procedure.

### 7.2.1 Participants

The participants were university lecturers from various institutes and disciplines. There were 12 participants in total consisting of 5 females and 7 males. The participant's ages, disciplines and locations are displayed in Figures 7.3a, 7.3b and 7.3c.

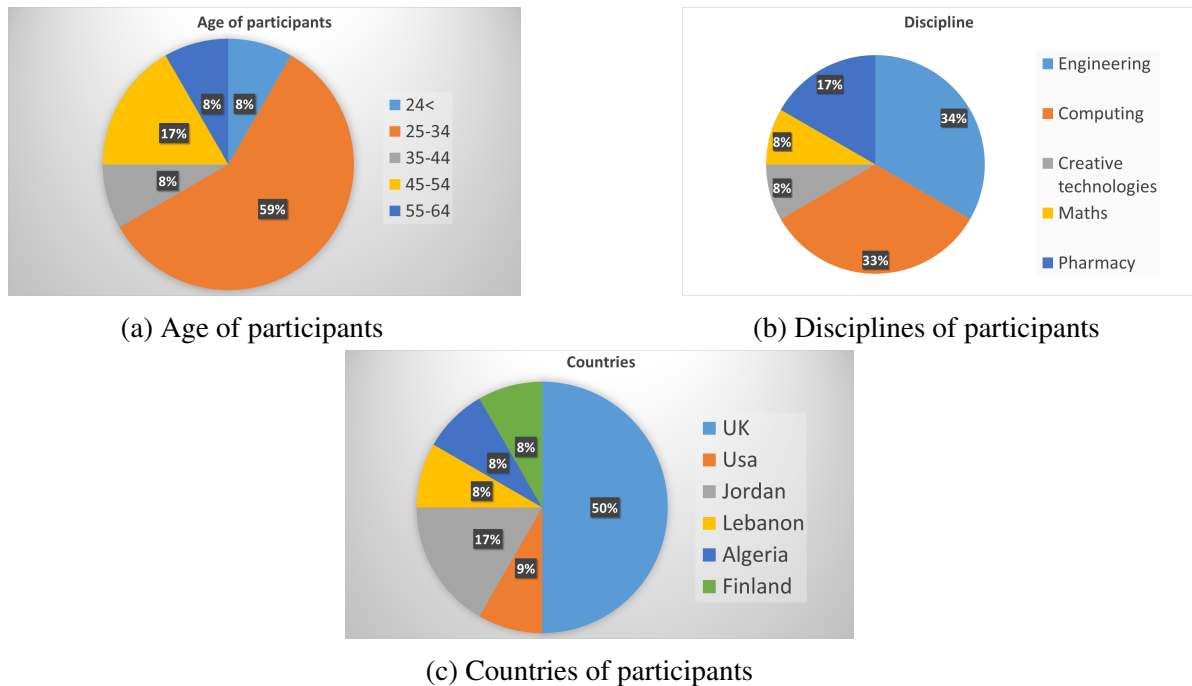


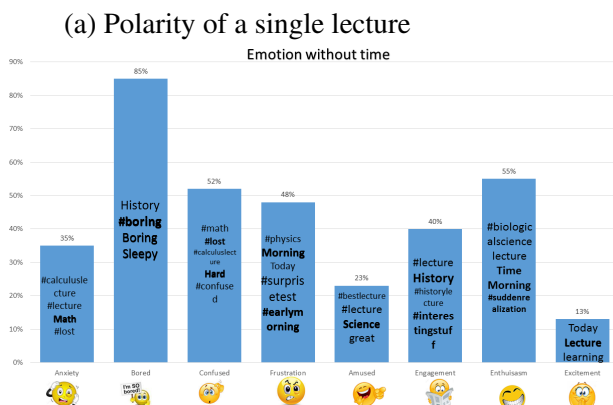
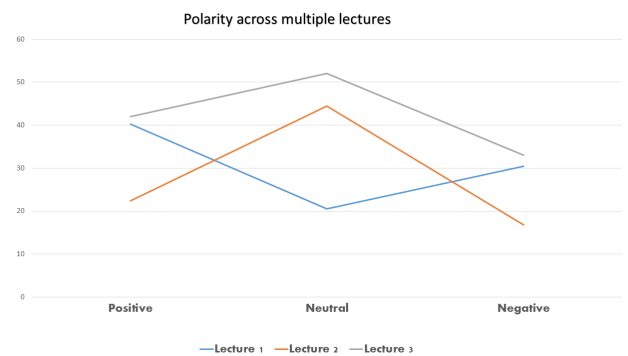
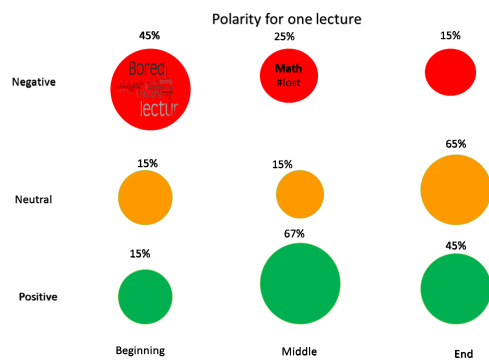
Fig. 7.3 Demographic information about the participants

### 7.2.2 Visualisation choices

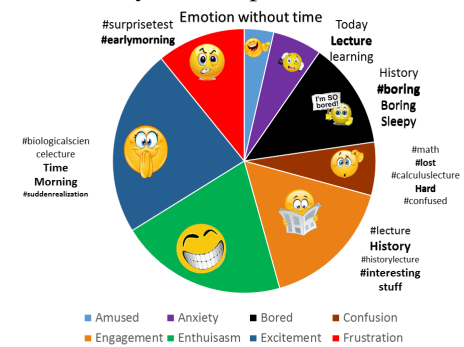
To find the best visualisations which are suitable for representing students' feedback, there is need of several visualisations including different choices. For instance, including visualisations with and without time. The default times provided were: Beginning (B), Middle (M) and End (E) of a lecture.

We created six visualisations, which are: Polarity with or without time, Polarity across different lectures, Emotions without time (2 visualisations) and Emotions with time (2 visualisations). The visualisations are presented in Figure 7.4.

The first visualisation was polarity for a single lecture with time. This visualisation was inspired by the bubble chart and is presented in Figure 7.4a. The axes were the time of

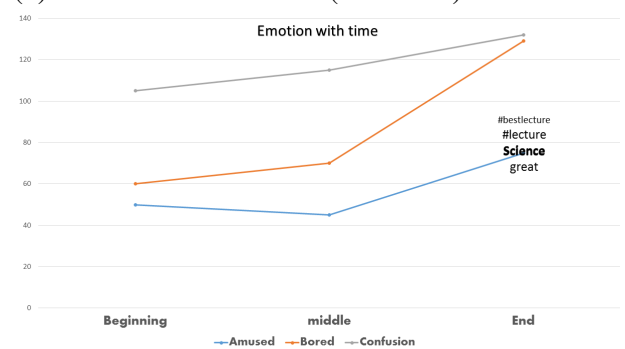
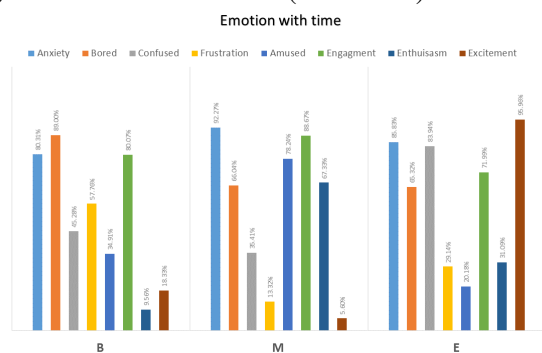


(b) Polarity of multiple lectures



(c) Emotions without time (Bar charts)

(d) Emotions without time (Pie charts)



(e) Emotions with time (Bar charts)

(f) Emotions with time (Line charts)

Fig. 7.4 All Visualisations

the lecture and polarity and the bubbles represented the percentage of students' feedback. The colours of the bubbles were red for negative feedback, orange for neutral and green for positive. These colours are generally associated with these polarities [105]. Words may make graphs more meaningful for the lecturer [110]. Some lecturers may like text appearance in visualisation, while others may not. Therefore, in the questionnaire, we presented lecturers with three options to display text (i.e. word list, wordle, no words at all). The words in the word list were at different sizes and some were in bold to show importance. Previous research also used word sizes and boldness to show word importance [13, 29, 39, 49].

The second visualisation was polarity across multiple lectures, which is presented in Figure 7.4b. This visualisation was based on a line chart type. We chose the line chart as it is easier to compare between multiple instances. Culbertson [37], Shah and Hoeffner [148] and Neri et al. [119] found that lines were useful to represent time. The axes of this visualisation were the polarity and its percentage and the lines represent the lectures. For each lecture, the feedback could be taken at different times in the lecture (i.e. beginning, middle and end). The average of the positive, negative and neutral from the different times were calculated. In the questionnaire, we asked lecturers if they would like to view the single distribution of the polarity for each of the different times.

The third and fourth visualisations were emotion percentages in the lecture without time of the lecture included. They are presented in Figures 7.4c and 7.4d. For these visualisations, we included bar and pie charts. A study Cleveland and McGill [33] showed that viewers might comprehend bar graphs more than pie charts. Similarly, Toth [172] found that pie charts present data poorly and bar charts were much more beneficial when presenting data. Lewandowsky and Spence [95] found similar results. In contrast, Schonlau et al. [145] claimed in their experiment that there was no difference between bar and pie charts in the viewers' responses.

We did not include stacked bar charts as Toth [172] claimed that it is difficult for the viewers to estimate the values presented and to make comparisons. We chose vertical bars over horizontal, as Culbertson [37] found that users prefer vertical bars, despite the fact that horizontal bars allow more space for labels.

We included emoticons in the graph to make it more presentable and to facilitate visualisation over time, when lecturers become familiar with the system and associate particular emoticons with particular emotions. This familiarity would allow the lecturers to identify particular emotions quicker via the emoticons, without the need to read the text. Additionally, frequent words were also included in both of the charts. In the questionnaire, different options were given to lecturers for the emotion charts. The first option was the bar chart order of bars: according to frequency or according to the negative and positive emotions (i.e. negative

emotions on the left and positive emotions on the right). The second option was the amount of emotions to be visualised: all the emotions or the most reliable emotions (according to our models results).

The fifth and six visualisations were students' emotions in the lecture with the time. These are presented in Figures 7.4e and 7.4f. Bar and line charts were used for these visualisations. A study by Shah and Hoeffner [148] highlighted that bar and line charts were useful when representing time. Shah and Hoeffner [148] highlighted that sometimes using lines to represent time is easier for the viewers to comprehend. On the other hand, Culbertson [37] found that participants preferred bar charts over line charts in his experiment. Zacks and Tversky [188] found similar results. The charts x-axis illustrated the time of the lecture and the lines and bars represented the emotion percentages.

### 7.2.3 Procedure

Two methods were used to collect lecturers' visualisation preferences depending on their location. The local lecturers were interviewed for around 15-30 minutes. The interview process involved presenting the visualisations to the lecturers. At the end of the interviews, lecturers were provided with a questionnaire. For the lecturers not present locally, the lecturers were provided with information about each visualisation in writing along with the same questionnaire.

### 7.2.4 Measurements

The questionnaire consisted of 20 questions. The first 4 questions were related to the demographic factors of the participants, including age, gender, disciplines and countries they are based in. One of the questions asked participants what devices they preferred to use to view the visualisation. There were some questions related to the options in the visualisations. They were asked for each of the visualisations to provide on a scale of 5 (strongly dislike to strongly like), how much they like the visualisation. The questionnaire is presented below.

1. *Q1: What is your age?*

- ☐ 24 or younger
- ☐ 25 to 34
- ☐ 35 to 44
- ☐ 45 to 54
- ☐ 55-64



- ☐ 65-74
  - ☐ 75 or older
- 2. *Q2: What is your gender*
  - ☐ Male
  - ☐ Female
- 3. *Q3: What department do you work in?*
- 4. *Q4: What country do you work in?*
- 5. *Q5: Which of the following devices do you most prefer to use to visualise students' feedback?*
  - ☐ Tablet
  - ☐ Desktop
  - ☐ Laptop
  - ☐ Smart phone
- 6. *Q6: How much do you like slide 1?*
  - ☐ Strongly like
  - ☐ Like
  - ☐ Neutral
  - ☐ Dislike
  - ☐ Strongly dislike
- 7. *Q7: In slide 1 what do you prefer to show words related to polarity (positive, negative and neutral)*
  - ☐ Wordle
  - ☐ List of words
  - ☐ None
- 8. *Q8: How much do you like slide 2?*
  - ☐ Strongly like
  - ☐ Like

- ☐ Neutral
- ☐ Dislike
- ☐ Strongly dislike

9. *Q9: In slide 2, would you prefer to see a detailed description of polarity for different times of the lectures?*

- ☐ Yes
- ☐ No
- ☐ Not sure

10. *Q10: In slide 1 and 2, would you prefer to see the neutral class, if it less reliable in accuracy?*

- ☐ Yes
- ☐ No
- ☐ Not sure

11. *Q11: For the emotion visualisations, how many emotions would you like to see?*

- ☐ All emotion
- ☐ Most frequent 3 emotions

12. *Q12: How much do you like slide 3?*

- ☐ Strongly like
- ☐ Like
- ☐ Neutral
- ☐ Dislike
- ☐ Strongly dislike

13. *Q13: In slide 3, how would you like the distribution of emotions to be (distribution of bars)?*

- ☐ Positive emotions and Negative emotions
- ☐ Emotions according to frequency

- 
14. *Q14: How much do you like slide 4?*
- ☐ Strongly like
  - ☐ Like
  - ☐ Neutral
  - ☐ Dislike
  - ☐ Strongly dislike
15. *Q15: How much do you like slide 5?*
- ☐ Strongly like
  - ☐ Like
  - ☐ Neutral
  - ☐ Dislike
  - ☐ Strongly dislike
16. *Q16: How much do you like slide 6?*
- ☐ Strongly like
  - ☐ Like
  - ☐ Neutral
  - ☐ Dislike
  - ☐ Strongly dislike
17. *Q17: In slide 3-6, some of the emotions are not reliable are you still interested in seeing them in the visualisation or just the reliable ones (Bored, amused, excitement)?*
- ☐ Yes, I would like to see all emotions
  - ☐ No, I only want to see the most reliable emotions (bored, amused, excitement)
18. *Q18: In slide 6, would you like to see the frequent words for each emotion when hovered over?*
- ☐ Yes
  - ☐ No
  - ☐ Not sure

19. *Q19: Is there anything that you would have liked to have seen in the slides that are missing?*
20. *Q20: Could you suggest a better visualisation, please take your time to draw one!.*

## 7.3 Results

The demographic graphics questions (1-4) were analysed and presented in the participants section 7.2.1. The participants were asked in question 5, what device they would prefer to use to interact with the system (i.e. view the visualisations). They were given four choices: desktop, laptop, tablet or smart phones. We found that the participants preferences were 58% laptops, 34% desktops, 8% tablets and 0% smart phones. This indicates that the application may be better being PC-based.

For each of the visualisations (Vis), we asked the participants how much they liked them on a scale of 1 to 5 (i.e. Strongly Dislike, Dislike, Neutral, Like and Strongly Like). These were covered in questions 6, 8, 12, 13, 14 and 15. The results are presented in Figure 7.5. To find the most liked visualisation, the participants scores for each visualisation scale were summed (See Table 7.2); for instance 3 participants strongly liked visualisation 1. Fisher's exact test was implemented to compare the visualisations. Fisher's exact test is a statistical test that compares between two nominal variables and finds if one variable values are different from the other. Fisher's exact test is suitable for small populations and was used in this study due to the low number of participants, i.e. 12. All possible combinations of 2 visualisations were compared to determine whether some visualisations were preferred to others. The p-value results of Fisher's exact test are presented in Table 7.3. The results show that none of the visualisation differences were significant, suggesting that the lecturers had no strong preference for particular visualisations.

Table 7.2 Visualisation score sums

Rate	Vis 1	Vis 2	Vis 3	Vis 4	Vis 5	Vis 6
Strongly Liked	3	2	1	4	4	3
Liked	3	7	8	5	3	7
Neutral	4	0	3	1	3	0
Dislike	1	2	0	2	2	2
Strongly Disliked	1	1	0	0	0	0

Fisher's exact test was used to compare between gender preferences in the visualisations. The results of these experiments are presented in Table 7.4. The results revealed that the

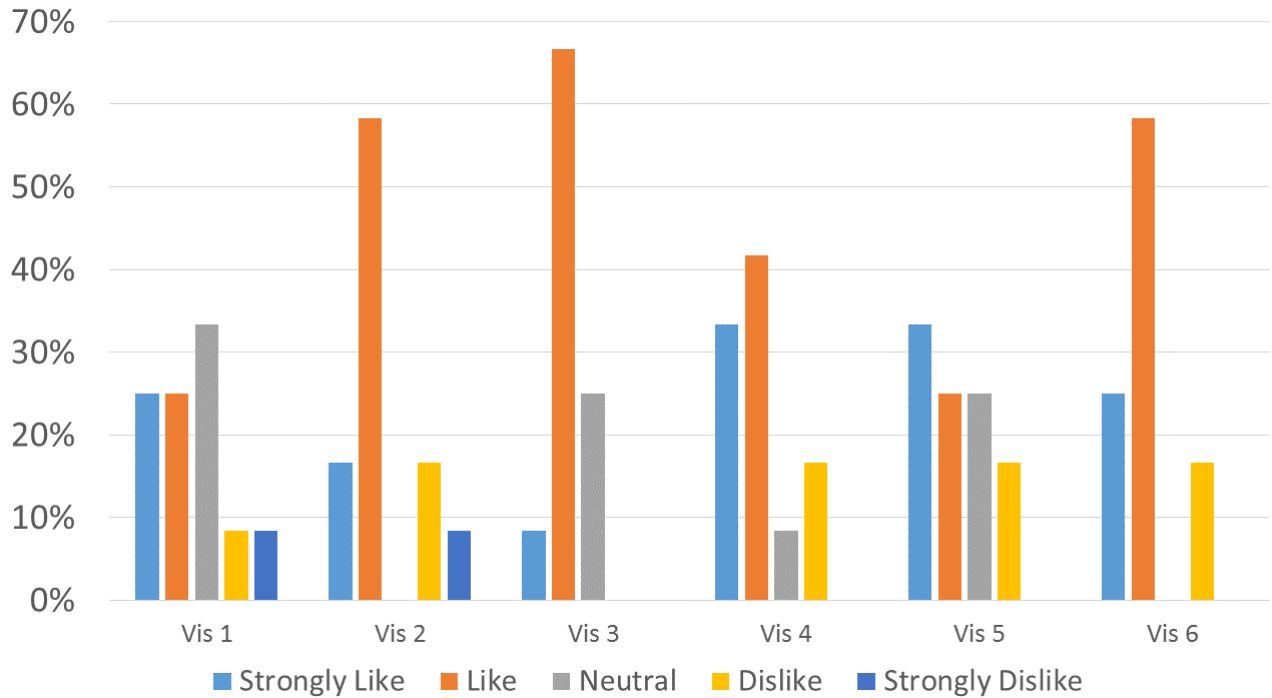


Fig. 7.5 Visualisation likeness

difference between males and females is significant in visualisation 5. Most of the males disliked the emotion visualisation with time in a bar chart form and preferred the visualisation with line charts, whereas, females preferred the emotion visualisation with time in a bar chart form.

As for the individual visualisations, we presented different options for the lecturers. In visualisation 1, the participants were asked if they preferred to view frequent words as wordle, word lists or no words at all (Question 7). The results showed that 33% preferred wordle, 42% word list and 25% no words. Most of the participants chose to view words over no words at all which confirms the previous literature findings in importance of words in sentiment analysis visualisation [88, 110].

The participants were also asked for visualisation 2, if they preferred to see detailed description of the polarity results (positive, negative and neutral percentages from all the different times in a single lecture), as the lines only visualise the average (Question 9). The results revealed that 83% liked to see this description while 17% were not sure. Some literature such as Gershon[60] claims that using too much information in the visualisation is confusing. However, in this case presenting details is preferred by lecturers to understand the visualisations components.

Table 7.3 Visualisation Likeness averages results

<b>Visualisations</b>	<b>P-value</b>
Vis 1xVis 2	0.1811
Vis 1xVis 3	0.2079
Vis 1xVis 4	0.5249
Vis 1xVis 5	1
Vis 1xVis 6	0.1192
Vis 2xVis 3	0.2391
Vis 2xVis 4	0.7315
Vis 2xVis 5	0.1794
Vis 2xVis 6	1
Vis 3xVis 4	0.1835
Vis 3xVis 5	0.09865
Vis 3xVis 6	0.1547
Vis 4xVis 5	0.8513
Vis 4xVis 6	0.8565
Vis 5xVis 6	0.234

Table 7.4 Gender Visualisation Likeness Comparison

<b>Visualisations</b>	<b>P-value</b>
Vis 1	0.07576
Vis 2	0.2576
Vis 3	1
Vis 4	0.4192
Vis 5	0.005051
Vis 6	1

For the polarity visualisations (i.e. visualisations 1 and 2), we asked participants if they preferred to view the neutral class if it was less reliable in accuracy (Question 10). This was based on our findings in chapters 4 and 5. Consequently, half of the participants preferred to view the neutral class, while 42% did not prefer to view it and 8% were not sure. It is interesting that only half of the participants chose to view the neutral class, while the literature in chapter 4 highlighted the importance of the neutral class in real-life applications. The preference could be influenced by the fact that the reliability of the models with the neutral class were lower. As for the emotion visualisations, we asked participants if they

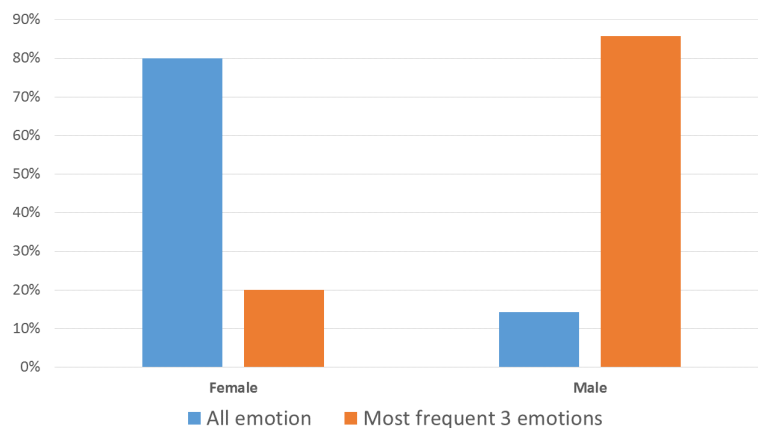


Fig. 7.6 Gender and number of emotions

preferred to view all the emotions or the 3 most reliable emotions (Questions 11 and 17). The results showed that 42% preferred to see all the emotions, while 58% preferred to see only 3 emotions. Additionally, as illustrated in Figure 7.6, we found that 90% males preferred to see only 3 emotions, while 80% of the females preferred to see all of the emotions. As for visualisation 3, which was the emotions without time with bar charts, we asked them which way they preferred to view the bars organised; according to frequency or the negative and positive emotions (Question 13). We found that 42% preferred to see all the bars according to the negative and positive emotions, while 58% preferred to see the bars according to frequency. The participants were asked if they prefer to view frequent words in the emotion line graph when hovered over the points (Question 18). The results revealed that 58% preferred to see frequent words, 33% preferred not to and around 1% were not sure. The results are similar and parallel to question 7, as participants preferred to view words.

The last two questions from the questionnaire were additional feedback and suggestions for new visualisations that could be considered (Questions 19 and 20). For the additional feedback, a few suggestions were given, for instance, one lecturer suggested that visualisation 2 could be used to compare between lectures from different subjects to rate lecturers'

performances. Another suggestion was to make the graphs automatic and to give lecturers trends in the students' feedback. This requires students to use their phones throughout the lecture, which is disliked in most lecturers' opinions. None of the lecturers suggested or drew a new visualisation.

## 7.4 Discussion

The results revealed that all the visualisation were liked by the lecturers. When comparing participant's scores for the visualisations, none were significantly preferred more than another. This suggests that various lecturers have different preferences; consequently, to make sure every lecturer would have access to their preferred visualisation, all of the proposed visualisations were implemented in the system. In contrast, the literature shows that some visualisations could be more favourable than others. For instance, Cleveland and McGill [33] found that participants in their study could comprehend bar graphs more than pie charts. Toth [172] also showed that pie charts present data poorly and are difficult for participants to understand and compare values and found that bar charts were more beneficial at presenting data.

We used Fisher's exact test to compare between gender preferences in the visualisations. The results of the Fisher's exact test on gender shows that females prefer visualisation 5 (bar charts), while male prefer visualisation 6 (line chart). Similarly to the female preference, Shah and Hoeffner [148] findings showed that line charts are preferred when presenting time. Likewise, Zacks and Tversky [188] found that viewers preferred bar charts over line charts. In contrast, Culbertson [37] claimed that participants preferred bar charts over line charts, which is similar to the male preference.

As for the polarity visualisations, we found that half of the participants preferred to view the neutral class. This could be due to lower reliability and accuracy of models within the neutral class. However, as shown in chapter 4, the neutral class is important to view a complete picture of the polarity and it does not force instances into positive and negative categories. Therefore, it is essential in the visualisation. Additionally, we found that participants preferred the word lists over wordle in presenting the word frequencies. This could be due to the simplicity of reading word lists in a small amount of time in the lecture. However, it is surprising that 25% of the participants did not prefer to see any words, as they usually give lecturers an indication of the reason for the polarity decision. In contrast, Mcnaught and Lam [110] found that word clouds were popular among viewers and useful to detect patterns in words. Culbertson [37] showed that word labels on graphs and pictorial symbols did not make any difference in participants' understanding. A possible reason for



some lecturers preferring to see no words is the use of these visualisations in real time. They would need to process the information from the visualisations while lecturing, and thus, would only be able to use a limited amount of attention for this purpose.

Our results showed that 58% of the participants preferred to view the order of the bars according to frequency in visualisation 3. This is easier for the lecturers to distinguish between the most frequent and less frequent emotions. Additionally, participants preferred to only view the 3 most accurate emotions which were amused, bored and excited. Our findings showed that 90% of the male participants preferred to only view 3 emotions, while 80% of the females preferred to see all the emotions. This could indicate that male only prefer simple visualisations over female who prefer detailed visualisations or that males prioritise accuracy, while females prioritise the ‘big picture’; however, this needs further investigation.

To conclude, we found that while some people prefer bar graphs, others may prefer line graphs or pie charts. The results showed that there is no clear preference for visualising the data and all the visualisations could be used. Thus, the best option is to present all visualisations.

One limitation of the research is the number of participants evaluating the visualisations. This was due to the difficulty in finding lecturers willing to participate. An email was sent out to different departments reaching over 100 lecturers, of which only 12 were willing to participate. This is a general issue and similar to other studies in the educational domain [6, 43, 108, 124, 149]. It could be due to the busy schedules of the lecturers especially during term time.

## 7.5 Summary and Contribution of the Chapter

This chapter reviewed different visualisations from previous research. Previous literature did not show any suitable visualisations for polarity and emotion results from students’ feedback. Six visualisations were created to present both polarity and emotions from students to the lecturer. We included different graphs and different options for the lecturer to choose from.

The methodology of this research included interviews and questionnaires to allow lecturers to participate and suggest different visualisation designs.

The results indicated that lecturers liked all the visualisations. In addition, the results suggested that all the visualisations were equally preferred as no significant value was found in Fisher’s exact test experiments.

The gender had an effect on visualisation 5 – we found that males disliked it, while females liked it. However, as the users of the system are a mix of gender, we can not exclude

this visualisation. However, this is interesting for future researchers to acknowledge the differences the gender have in this particular visualisation.

The contribution of this chapter was to create visualisations suitable to present the results of the polarity and emotion results from students' real-time feedback to lecturers. Our findings revealed that while some people prefer bar graphs, others may prefer line graphs or pie charts. Therefore, we found that the best option is to present all visualisations. Another contribution was to explore the effect of gender in the visualisations. Our results showed that there is a relation between gender and types of visualisation, which is interesting to explore further in future research.

# Chapter 8

## System integration and evaluation

The sentiment analysis models and the visualisations were integrated into a PC-based system to be used in real lectures. The system architecture will be described in section 8.1, followed by the system evaluation in section 8.2. The chapter is ended with a discussion and conclusion in sections 8.4 and 8.5.

### 8.1 System Architecture

The system consists of the sentiment analysis models which predict polarity and emotion from students' feedback and visualisations that present the results to the lecturer. Figure 8.1 illustrates the components of the system and Figure 8.2 illustrates an example of its process. The system was built and integrated using R language.

One scenario that illustrates the benefits of the proposed system is described in the following. John, a lecturer, has just finished presenting an example, and he wants to know whether to continue to the next part of the lecture. The students are instructed to post tweets using their own mobiles, tablets, or laptops to express their opinions about the lecture. The feedback is then analysed using the sentiment analysis models and is visualised to the lecturer. The visualisation illustrates different proportions of positive, negative and neutral sentiment. John can see that the percentage of negative feedback is 70 percent, the neutral is 20 percent, and the positive is 10 percent. The visualisation also presents frequent words with their polarity; for example, the words 'confused', 'example', 'complicated' and 'difficult' appear with negative polarity. As a result, he can address this in real time and decide to explain the example in a different way.

The system will be implemented in the classroom as follows: Students will provide feedback to the lecturer via social networks such as Twitter. Social networks are popular among students [123, 174]. In 2010, Twitter a social media network, had 106 million users,

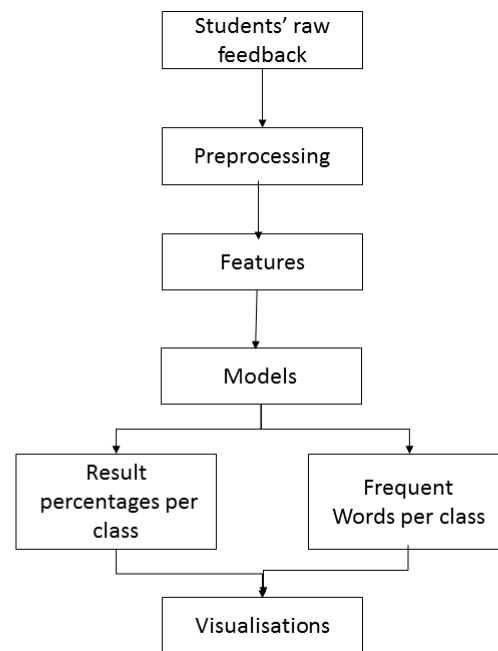


Fig. 8.1 System components

and 180 million visitors every month [15]. Twitter company revealed that 300,000 new users were signing up every day and that it received 600 million queries daily through its search engine. Twitter is also free, and around 37% of Twitter's users use their phone to send messages [15]. Twitter is also popular in sentiment analysis research as it is rich with opinions [1, 12, 63, 64, 92, 126, 136, 181]. Therefore, Twitter can be used by students to provide feedback. Using Twitter allows students to express feedback in their own wording, which allows the lecturer to view frequent words per polarity/emotion class. This can be an advantage over clickers that only retrieve limited feedback such as Yes/No.

The students' feedback can be taken at anytime in the lecture but there will also be time slots chosen by the lecturer according to stop points, such as changing topics or after complex examples. For our system we choose 3 time slots which are beginning, middle and end of the lecture. This distribution is beneficial as the feedback might vary at different times in the lecture.

The feedback is then passed through the sentiment analysis models. This part consists of three stages: preprocessing the data, selecting the features, and passing through the model which is based on a particular machine learning technique. A description of the optimal models labelling the feedback have previously been mentioned in chapters 4 and 5. Both of these models include a low preprocessing level including tokenization, converting text to

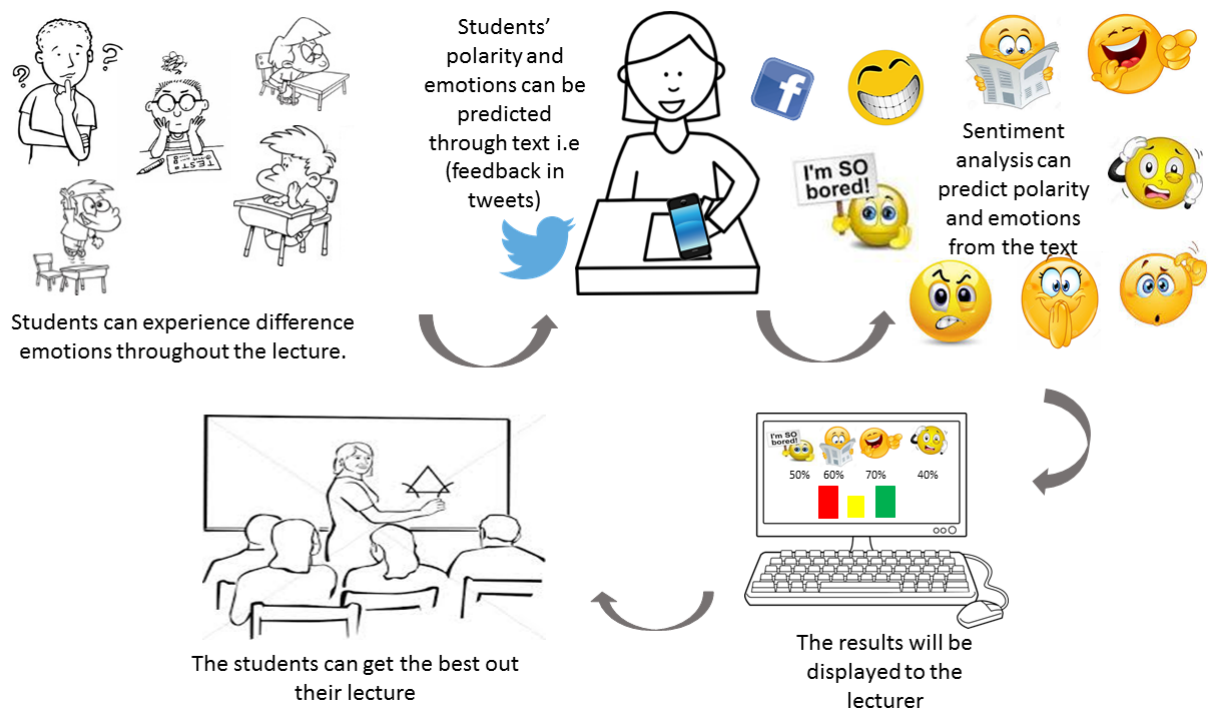


Fig. 8.2 System process

lowercase and removing numbers, punctuation and stop-words. They both include unigrams as features and CNB as the machine learning technique. The polarity models included the neutral class. The emotion models included three single emotion models (i.e. Amused, Bored and Excited). The results are associated with frequent words from each class in the polarity and emotions using the term frequency function in the text mining package in R.

The results from the models are fed into the visualisations in the system. According to Chapter 7 findings, there were no strong preferences with regard to the 6 visualisations. The lecturer can view their preferred visualisations from the 6 available ones, then take actions accordingly.

## 8.2 System evaluation experiment

In this section, the system evaluation is presented. The purpose of evaluating the system is to investigate if it is useful and usable. Additionally, to find whether or not students and lecturers adapt to the system in real lectures. The participants and the details of experiments procedure are highlighted, followed by the results.

### 8.2.1 Participants

The participants consisted of 80 students and 6 lecturers from the computer department at an U.S. Institute, which cannot be named for ethical and legal purposes (i.e. they requested to remain anonymous). The number of student participants were 80. Some of the students in the class did not participate in the experiment, due to their unwillingness to fill in the questionnaire. Details about the lectures included in this experiment and the number of students that participated in each class are presented in Table 8.1. Some of lectures were selected from the same course to observe if there is any difference in the lecturer/students opinions.

Table 8.1 Lecture Details

Date	Topic	Year	Students in class	Students participated
14-Sep	Advanced Object-oriented Programming Using C#	Year 2	12	12
22-Sep	Introduction to Operating System	Year 1	15	15
28-Sep	Server-side Scripting	Year 2	19	19
28-Sep	Computer Information System (CIS) Project	Year 3	14	13
30-Sep	Server-side Scripting	Year 2	19	16
30-Sep	Computer Information System Project	Year 3	10	5

### 8.2.2 Procedure

The purpose of this study was explained to the students by their lecturers. The students were asked to provide feedback via Twitter at different times of the lecture. They were provided with a Twitter account to tweet or they could use their own accounts. The lecturers of this experiment interacted with the system. At the end of each lecture, the students and lecturer were asked to complete two separate questionnaires.

The students were asked a short questionnaire consisting of 4 questions. It included questions about if they participated or not and if they liked/adapted the system process in the lecture.

The lecturer questionnaire was denser, as they were the users of the system. The questionnaire included questions to find how useful and usable the system was to them. It also included questions to see if they adapted to the use of the system in their lectures. The lecturer questionnaire consisted of 15 questions. Some of these questions were based on

previous popular questionnaires from the literature to evaluate systems [96, 98, 103]. Details about these questionnaires were previously discussed in chapter 3 in section 3.3.3.

### 8.2.3 Measurements

The students' questionnaire consisted of four questions. The first question was a Yes/No question asking students if they participated in providing feedback. The three remaining questions were Likert type questions with a scale of 5 options (i.e. Strongly agree to Strongly disagree):

- *Q2: I feel that the lecturer changed in the lecture according to my feedback.*
- *Q3: I like this new teaching style.*
- *Q4: I think the system is useful for me (i.e. providing feedback which will be responded to in real time)*

The lecturer questionnaire consisted of 13 Likert questions:

- *Q1: Overall, I am satisfied with the use of the system.*
- *Q2: The system was easy to use.*
- *Q3: The visualisations provided are easy to understand.*
- *Q4: I feel that the system gave an accurate and correct judgement about the classes' opinions.*
- *Q5: I can integrate the system into my everyday teaching.*
- *Q6: The system helps you to understand students more.*
- *Q7: The system in general is a distraction in class.*
- *Q8: I found the polarity detection useful.*
- *Q9: I found the system useful.*
- *Q10: Students using their mobiles are a distraction in class.*
- *Q11: I found the emotions detection useful.*
- *Q12: I found the visualisations useful.*

- *Q13: I changed my teaching based on the systems results.*

We also asked lecturers how often they used the system in one lecture (“beginning, middle and end”, “only end” or “all throughout the lecture”). Lastly, we asked lecturers if they found any difficulties using the system.

## 8.3 Results

The questionnaires were evaluated in different methods according to the number of participants. The students’ questionnaires were evaluated using parametric statistical tests. The single lectures part of the students’ questionnaire and the lecturer questionnaires were evaluated using nonparametric tests due to the small sample size. Nonparametric tests are calculated by observing the median from a sample, while the parametric ones are calculated by observing the mean. In the following subsections the student and lecturer questionnaire results are presented.

### 8.3.1 Student Questionnaire Results

The first question explored if students participated in providing feedback or not. The results show that one student answered ‘No’ to this question and two students left the answer blank.

Cronbach’s alpha test was used to test reliability of the answers of the three questions. Cronbach’s alpha is a test used to measure scale reliability and internal consistency between items (i.e. questions). Cronbach’s alpha depends and is correlated with the number of cases (i.e. participants in the study). The formula of Cronbach’s alpha is:  $\frac{N*c}{v+(N-1)*c}$ , where  $N$  is the number of items,  $c$  is the average inter-item covariance among the items and  $v$  equals the average variance. Cronbach’s alpha is a value between 0 and 1. The acceptable ranges of alpha are from 0.70 to 0.95 [161]. The result of Cronbach’s alpha on our data was 0.918. This indicates that the results are reliable.

One sample t-tests were also used for each question. One sample t-tests are used to compare the mean of a sample to a particular value. The average mean value of the three questions was calculated which we name student question average. The student question average was used to explore the overall mean across the three questions. To establish if the students liked or disliked the use of the system, the means of the single questions and the student question average were compared with the middle value of the Likert scale, i.e. 3, which indicates a neutral position. The results of the student one sample t-tests are presented in Table 8.2. The only difference that was not significant was in Q4 (i.e. the lecturer changes the style of teaching). The student question average mean value was 2.6917 ( $t(79)=-2276$ ,



Table 8.2 Lecture Details

Sample	Mean Value	P-Value
Q2	2.700	0.025
Q3	2.675	0.033
Q4	2.700	0.059
Composite Average	2.6917	0.026

p-value= 0.026), which indicates that the results are significantly lower than the middle value 3. Therefore, this suggests that the students have a relatively negative impression of the system.

The number of lectures in the experiment were 6. We investigated the individual lectures to find if students liked/disliked the use of the system in some lectures and not in others. For this experiment One sample Wilcoxon Signed Rank Test was used, due to the small amount of data distribution in each lecture. Similar to the previous experiment, the value 3 was taken for the sample comparison. The results of tests are presented in Table 8.3. We found that only some of the results differences were significant. Lecture 2 and Q2 from Lecture 4 were the only values that were significantly lower than the set median 3. In general, by observing the median results, we notice that the lowest results occurred in lectures 2, 4 and 6. Lecture 2 was introduction to programming. The low results could be due to the nature of the topic being a theoretical topic and different from other lectures which are more practical. In addition, they may have not seen any use in the system as lecturers did not make any change in teaching according to their feedback.

The highest median rate across all lectures was for those including year 2 students (Lectures 1, 3 and 5). This could suggest that the system is more useful and acceptable in lectures with students from this range (i.e. year 2).

Lectures 3 and 5 were from the same topic (i.e. Server-side scripting). They both had a median rate of 3, which is a higher mean rate in comparison with the other lectures. Lectures 4 and 6 were also from the same topic (i.e. CIS projects), both had a median rate of 2. This could be due to the nature of this course being at an advanced level (i.e. year 3) and there is less interaction with the lecturer.

To compare between the lectures with the same topic, the means of the question were calculated (See Table 8.4). As we can notice from the results, the means for lecture 5 were higher than in lecture 3, suggesting that students liked the use of the system better in the second class. On the other hand, in lectures 4 and 6, we found that mean value drops in the second lecture, indicating that students did not like the use of the system as much as in

Table 8.3 One sample Wilcoxon Signed Rank Test for individual lectures

Lecture	Sample	Median	P-value
Lecture 1	Q2	3	0.276
	Q3	3	0.420
	Q4	3	0.292
	Composite Avg.	3	0.442
Lecture 2	Q2	2	<b>0.015</b>
	Q3	2	<b>0.003</b>
	Q4	2	<b>0.015</b>
	Composite Avg.	2	<b>0.003</b>
Lecture 3	Q2	3	0.902
	Q3	3	0.495
	Q4	3	0.371
	Composite Avg.	3	0.621
Lecture 4	Q2	2	<b>0.023</b>
	Q3	2	0.090
	Q4	2	0.158
	Composite Avg.	2	0.052
Lecture 5	Q2	3	0.713
	Q3	3	0.584
	Q4	3	0.301
	Composite Avg.	3	0.304
Lecture 6	Q2	2	0.129
	Q3	2	0.102
	Q4	2	0.063
	Composite Avg.	2	0.078

the first class. These results are conflicting therefore it is difficult to make a judgement to whether students will like or dislike it more in time.

### 8.3.2 Lecturer Questionnaire Results

There were 6 participants in the lecturer study, therefore it is assumed that the data is not normally distributed. One-Sample Wilcoxon Signed Rank Test was used due to it being an equivalent to One sample t-test. The One-Sample Wilcoxon Signed Rank procedure assumes that the sample we have is randomly taken from a population.

There were 13 Likert questions. The median for all the questions ranged from 4 to 5, indicating that the system was highly liked and that the lecturers were satisfied. The average median over all questions was 4. The One-Sample Wilcoxon Signed Rank test was applied to all the questions with the median value of 3 (the neutral value). The results of the tests

Table 8.4 Mean for same topic lectures

Lecture	Server-side scripting		CIS project	
Sample/Lecture #	Lecture 3	Lecture 5	Lecture 4	Lecture 6
Q2	2.95	3.125	2.154	2
Q3	2.85	3.188	2.308	2
Q4	2.7	3.375	2.462	1.8

are presented in Table 8.5. The results revealed that nearly all questions were significantly higher than the middle value 3 with a p value  $<0.05$ . The only two questions which were not significant were: the system being a distraction in class and changing the teaching based on the system results. This is similar to the findings of the students' questionnaire in Q4 p-value. The lecturers were also asked if they used the system all throughout the lecture (i.e. allowing

Table 8.5 One-Sample Wilcoxon Signed Rank test for individual lectures

Sample	Median	Mean	P-value
Overall, I am satisfied with the use of the system.	4	4.17	0.02
The system was easy to use.	5	4.83	0.02
The visualisations provided are easy to understand.	4	4.33	0.023
I feel that the system gave an accurate and correct judgement about the classes opinions.	4	4.33	0.023
I can integrate the system into my everyday teaching.	4	4.00	0.014
The system helps you to understand students more.	5	4.83	0.02
The system in general is not a distraction in class.	4	3.50	0.18
I found the polarity detection useful.	4	4.33	0.023
I found the system useful.	5	4.83	0.02
Students using their mobiles are not a distraction in class.	4	4.16	0.02
I found the emotions detection useful.	4.5	4.50	0.024
I found the visualisations useful.	4.5	4.50	0.024
I changed my teaching based on the systems results	4	3.67	0.102

students to tweet at any time), at the beginning, middle and end of the lecture, or only at the end of the lecture. We found that 17% used the system at the beginning, middle and end of the lecture and 83% chose to use the system all throughout the lecture. This is interesting as some of the previous research indicated that social media and using mobiles may be a distraction to the students. Additionally, the lecturers have mostly claimed that the system for the students was not a distraction. This is a good indicator that the system can be used in real life. Lastly, the lecturers were asked if any part of system was difficult to use and 50% answered no, while the rest did not include any answer.

## 8.4 Discussion

The results of the students' questionnaire suggest that the students were willing to participate in providing feedback (80% of the students participating). This is an important and essential component in the system as without the students' willingness to participate, there would be no feedback to analyse. Similar to other systems used to collect feedback such as clickers [117, 134], mobiles [146, 154] and social media [121, 139], we found that students are willing to provide feedback in class. This can suggest that our system allows an increased interaction between the lecturer and the students. Consequently, according to research by Poulos et al. [135] the increased interaction could lead to improvement in teaching.

The students were asked to rate the system in three questions in a score of 1 to 5. The means of the questions ranged from 2.65 to 2.70, which is between the "disagree" and "neutral" option, being slightly closer to "neutral"; this indicates that overall, the students have a slightly negative view of the system. We found from both the questionnaires that the lecturer did not change their style in teaching according to the feedback. Additionally, students' opinions about the system usefulness was close to neutral. This could be due to them not finding any value or importance in providing feedback due to the lecturers not changing according to their feedback as both students and lecturer questionnaires indicate that no change was made to the lecture. On the other hand, students provided feedback which is a very valuable component of the systems success. Therefore, the possible explanation of the slightly negative views could be due to the lecturers not applying any changes to their teaching.

A comparison of lectures (i.e. using the mean value) with the same topics was done to investigate whether the opinions of students changed over time. The mean values suggest that there is no clear answer to whether students like the system better or worse due to conflicting mean values. The results suggest that there is a relation between the mean value in Q2 and Q3 and Q4. When Q2 is lower, Q3 and Q4 are also lower suggesting that students low feedback could be due to the lecturer not changing their teaching according to students' feedback.

The lecturers' evaluation of the system were positive, with the average median value of the lecturers responses over all the questions resulting 4 out of 5. This suggests that lecturers were satisfied with the system. Almost all of the questions were significantly higher than the median average, indicating that the lecturers were satisfied with the system. Only two questions were non significant: the system being a distraction in class and changing the teaching based on the system results. This indicates that lecturers welcomed students' feedback but preferred not to change their teaching according to it.

The system was found useful and usable by lecturers, which is similar to other SRS studies [117, 134]. Lecturers strongly felt that the system was easy to use and the system was

useful (Mean=5, p-value=0.02). The results also indicate that the system helped lecturers understand students better (Mean=5, p-value=0.02). Similarly, previous literature also reveals that students' feedback, allows the lectures to understand their students more [24].

In comparison to other studies who have collected students' feedback in real time via Twitter such as Novak et al. [121], we found that the lecturers did not find any difficulties or stress in using the system. Novak et al. [121] mentioned that lecturers had to read tweets from the beginning in sequential order causing time-loss and stress.

In comparison to other systems used to collect feedback in real time using SRSs, we found that our system has an advantage over other systems. Our system overcomes certain challenges found in SRS, for instance, lack of information issue in clickers which was addressed by including frequent words. Additionally, by using social media, the additional costs found in the purchase of clickers and SMS to provide feedback are eliminated [86, 134].

Limitations of this chapter include the small number of lecturers evaluating the system. However, in comparison with other research in education this is a common problem, due to the busy schedules of lecturers and students. Table 8.6 presents some examples of previous research indicating that collecting participants in the educational domain is difficult and a general problem.

Table 8.6 One-Sample Wilcoxon Signed Rank test for individual lectures

Paper	Participants	Topic	# of participants
Ozkan [124]	Students	Students' evaluation of e-learning systems in the higher education context	84
Martyn [108]	Students	Clickers in the classroom	184
Shahriza et al. [149]	Students	Mobile phone in collecting students' feedback	240
Anderson et al. [6]	Students	PC Based Lecture Presentation System	477
Davis et al. [43]	Students	Massively Open Online Courses	258

Another limitation to the study was that it was only implemented on lecturers from the computing department. The results may be distorted for this reason and it would be interesting to apply this study to lecturers in other departments.

## 8.5 Summary and Contribution of the Chapter

In this chapter, we discussed the system evaluation. The system was evaluated in 6 different lectures. The students and lecturers both participated in providing feedback about the system using questionnaires.

The results of the students' questionnaires suggested that students' opinions of the system were between neutral and dislike. This could be due to the lack of changes applied to the lecture after providing feedback. However, students were willing to participate in providing feedback, which is a prerequisite for the system to provide information to the lecturer.

The results of the lecturers' questionnaires indicated that the lecturers were satisfied with the use of the system. Across all the questions the median of the results ranged from 4 to 5 out of 5, which suggests that lecturers found the system useful and easy to use. They also found that the system allows them to understand their students better. On the other hand, we found that lecturers did not apply changes in the lecture according to students' feedback.

The contribution of the chapter was the integration of the system and the evaluation its use in real-life settings. The system was evaluated in multiple lectures with students from different levels.

## **Chapter 9**

# **Summary, Contribution and Future Work**

This thesis explored the use of sentiment analysis in students' real-time feedback. The significance of this research is that it addresses several gaps in the sentiment analysis literature. The research is important in the areas of polarity, emotion and sarcasm detection, visualisation and technology enhanced learning. This thesis explored polarity, emotion and sarcasm detection in students' feedback. In addition, visualisation of the polarity and emotion detection results was also investigated. The system proposed in this thesis which consisted of the integration of the models' results and their visualisations was evaluated, contributing to the field of technology enhanced learning.

### **9.1 Summary and Contribution**

In chapter 2, the main literature related to this research is reviewed and several gaps were identified. Previous methods and systems to collect feedback in real time were reviewed. These systems have certain limitations, such as the fact that the clickers and SMS messages were expensive for the institute and students. Additionally, clickers provided the lecturers with limited information due to their restricted capability. The most suitable method found to collect feedback was from social media websites. Social media websites are free and eliminated the cost charges from both the clickers and SMSs. In addition, social media websites were found popular among students. The use of social media, however, poses a different problem: how to analyse the collected feedback without stress and time pressure for the lecturers?

In result to the literature findings, we proposed the creation of a system that analyses students' feedback by detecting polarity and emotion and visualises the results to the lecturer in real time. We identified sentiment analysis as the most suitable method to analyse the feedback, due to its similarity to previous feedback collection methods such as clicker application and its ability to extract useful information from the text.

Previous research related to sentiment analysis was reviewed. The literature showed that sentiment analysis was more effective when applied to specific domains due to the different word meanings across domains. Previous literature related to sentiment analysis in different domains including education were reviewed. Most of the research in the educational domain was on e-learning and only one piece of research was applied to classroom feedback. The extensive literature review did not reveal any research exploring sentiment analysis for students' real-time feedback. In addition, the literature findings showed that there is no model fit for analysing students' classroom feedback, which is the main reason for this research.

Research related to emotion detection was also reviewed. The literature showed that there is little research related to emotion detection via text. Previous research highlighted the importance of emotion in learning. The literature indicated that there is no research that had explored emotion detection in students' classroom feedback. Additionally, to the best of our knowledge, no research has been done in an educational context on text data using machine learning, with the exception of one study on Chinese text.

The first contribution of this thesis was to explore polarity detection in students' feedback. This contribution was outlined in chapter 4. Two experiments were done: the investigation of n-gram and POS-tagging and the investigation of other features.

The first experiment included exploring different preprocessing levels with the most common features which were n-grams and POS-tagging and the most popular machine learning techniques. Additionally in this experiment the neutral class was also explored. One contribution to this study was to investigate the effect of preprocessing on polarity. Consequently, after exploring 6 different preprocessing levels, we found that preprocessing increased the performance of the models. The second contribution was to explore the common features in the literature (n-grams and POS-tagging). Our results showed that n-grams led to a better performance than POS-tagging for our experiment. Another contribution was to identify the best machine learning algorithms for detecting polarity in students' feedback. As a result, we found that LIN, NB and CNB had the highest performances. Additionally, the neutral class in our application was found important and the CNB classifier performed well with models with the neutral class.

In the second experiment, additional features were explored which were lexicon-based, tweet-related, general text-related and domain-based features. Three additional classifiers (i.e.



MNB, SMO and RF) were investigated. We also explored with methods to address unbalanced data. In opposition to the first experiment, we found that preprocessing the data lowered the performance. Based on our results, the additional features increased the performance of the models compared with unigrams alone. On the other hand, some of the features may be difficult to obtain in real-life settings. Similar to the first experiment, the CNB performance was high and it could handle the unbalanced classes. To further investigate addressing unbalanced classes, undersampling and oversampling were explored. Our results showed that undersampling did not have any impact on our data, while oversampling increased the performance slightly. Despite the usefulness of oversampling, CNB achieved a good performance without it.

One of the main contributions and aims of chapter 4 was to identify the best performing model for polarity detection in students' feedback. This included finding the best combination of preprocessing techniques, features and machine learning techniques. The optimal model found consisted of a lower level preprocessing (i.e. converting text to small case and removing numbers, punctuation and stop words), unigrams for features and CNB for the classifier.

The second contribution is to explore emotion detection in students' textual feedback. This contribution was presented in chapter 5. Two experiments were presented: exploring the n-grams and exploring other features.

In the first experiment, we explored preprocessing in emotion detection and found that it leads to a worse performance due to it removing valuable elements for the detection. We explored different models which detect multiple emotions or single emotions. We found that models which detect a single emotion have a better performance than multiple emotion models. Additionally, the best performing models found were three single emotion models (i.e. Amused, Bored and Excited). One of the contributions for this study was identifying the features that lead to the best performance. Our results showed that unigrams and trigrams led to the best performance for the 2-Class models, while unigrams combined with bigrams and trigrams led to the best performance for the multi-class models.

In our second experiment, we built on experiment 1 and explored additional features which were: domain-related features (i.e. lexicon-derived and time of the lecture), sarcasm, Twitter-related features (i.e. number of hashtags and emoticons), punctuation-related features (i.e. number of punctuations signs and question and exclamation marks) in combination with unigrams. We explored different combination of these features according to their ranking in a feature usefulness experiment. As a result, these features improved the results of the models in comparison with the unigrams alone. We found that using 15 additional features led to the highest performance in the majority of models. One limitation to experiment 2 is that applying some of these features (i.e. sarcasm) in real-life settings is unrealistic due to them

requiring additional information from the users (i.e. students). Another contribution of this study was to identify the best machine learning algorithm to detect emotions. Consequently, similar to the polarity detection, we found the best classifier across both of the experiments is CNB. MNB, ME and SMO also performed well, however, CNB addresses the imbalanced dataset therefore was selected for the application.

To summarise, the main contribution of chapter 5 was to identify the best models for emotion detection in students' feedback. The optimal model included single emotion detection models. It included a low level preprocessing method, unigrams as features and CNB as the classifier. Unigrams without any additional features were chosen as a feature. We justified this choice as in sentiment analysis, it is better to use the least number of features to obtain the results due to the lower complexity of the model which has an implication on the processing time.

One of the main limitations of sentiment analysis is the presence of sarcasm and the distorting effect it can have on the results. Various studies have explored sarcasm in sentiment analysis. However, there was no research exploring sarcasm detection in students' feedback which was the third contribution of this thesis. The contribution was presented in chapter 6.

The main aim of the study in chapter 6 was to identify the best models to detect sarcasm from students' feedback. In this study, two methods were used to label students' sarcasm: mismatched labels and detecting 'sarcasm' and 'sarcastic' words and hashtags. Two classes were used: sarcastic and non-sarcastic. One contribution was to explore the best preprocessing level for the detection. Three preprocessing levels were investigated and we found that the lowest preprocessing level (P1) led to the highest performance. Another contribution was to identify the features that lead to the best performance. Unigrams combined with 7 other features were explored. Our results showed that by adding emotion and polarity labels, sarcasm detection increased by 1 to 6%. Additionally by including all the features with unigrams, there was an increase in sarcasm detection in the NB, LIN and SMO classifiers by 7%, 4% and 9% respectively. Similar to our previous experiments, we explored several classifiers. We found the best classifier was CNB which had a recall of 63%. On the other hand, RB, SMO and RF classifiers had the lowest recall indicating they were not suitable for sarcasm detection on our data.

The fourth contribution of this thesis was to explore visualisation of sentiment analysis models' results. The main aim was to visualise the results of our models to the lecturers. In chapter 7, we explored previous research related to visualisation and particularly in visualising sentiment analysis. We created six visualisations to present both polarity and emotions from students in the lecturer. Different graphs and options were provided to the lecturer.

The results of chapter 7 showed that lecturers did not have a preference in any of the visualisations. The results indicated that some people prefer bar graphs, others may prefer line graphs or pie charts. Consequently, we found that the best option was to present all visualisations. We also explored gender's impact on the visualisations and found that it had an effect on visualisation 5 (i.e. males disliked it, while females liked it). This is interesting for future researchers to acknowledge the differences the gender has in this particular visualisation.

The last contribution of this thesis was to investigate the systems usefulness and usability in lectures. In chapter 8, we presented the system integration and evaluation. The system was evaluated in real-life settings which included 6 different lectures. The evaluation included students and lecturers providing feedback about the system using questionnaires.

The results of the students' questionnaires indicated that students' opinions of the system were between neutral and dislike. The students were willing to participate in providing feedback, which is an important component in the success of the system. The results of the lecturers' questionnaires indicated that the lecturers were satisfied with the use of the system. The median of the question results ranged from 4 to 5 out of 5. Moreover, the results showed that the lecturers found the system useful and easy to use. The lecturers also found that the system allowed them to understand their students more. However, our results show that lecturers did not aim to change their teaching according to students' feedback, which may have been the main cause for students disliking it.

## 9.2 Limitations

Like any research, the work presented in this theses has some limitations. One of the challenges across all the experiments in this thesis was collecting feedback. This is a general challenge in the educational domain and could be due to the busy schedules of faculty members and students.

In prediction models, one cannot cover every possibility, however, there are several ways of compensating for this. Having a limited amount of data only allows us to cover some of the cases, therefore, in sentiment analysis it is preferable to have a large amount of data. Limitations to sentiment analysis in general include that we can not know what each individual means in their textual feedback. Sentiment analysis can predict the possibility that ones sentence may reflect positive, negative or neutral however there is a chance that the model predicts incorrectly; however, it is also possible for people to misinterpret the polarity of a sentence.

In emotion detection, it is even more difficult to give a correct prediction due to the numerous emotions one can have and the closeness of emotions. An example for this in the educational domain is the emotions: Excited and Enthused. To be able to get an accurate prediction, the model needs a large amount of data for each class. This is probably the reason why models that detect multiple instances perform worse than models that detect a single emotion.

Another limitation is not knowing the cultural backgrounds of some people. The universities in this study have a high number of international students. We do not know if the cultural differences have any effect in the way students expressed themselves in the feedback they provided. Culture differences could also have an effect on the students' use of sarcasm; for instance, sarcasm may be more often used in the British culture compared with other cultures [35].

Although this research is based on English language, English language can vary and each culture has their own impact on it. There exist many differences between English vocabularies in the UK and the USA [36]. On the other hand, most of the word differences are caused from slang (informal language) and based on our findings the language and terms used in students' feedback were in standard English (formal language). The terms were related to education such as the term 'lecture' which can be used across different cultures.

Another issue with sentiment analysis is the evolution of language and the addition of new terms in the vocabulary over time [2]. The reasons for language change are diverse, and include the introduction of new terms for referring to new technologies [2]. Therefore, the sentiment analysis models need to be updated over time to be suitable for their purpose.

One limitation from the polarity, emotion and sarcasm experiments with other features is the implementation of the models in real-life settings due to the difficulty in obtaining some of the features without the requirement of additional information. Therefore they can not be applied into the system.

Another limitation was in the evaluation of the system, as it was only evaluated in the computing department. Additionally, we did not explore gender effect in the evaluation of the system due to the difficulty in collecting data and the additional challenges for balancing the data according to gender.

### 9.3 Reflections

In this thesis, we covered different sentiment analysis models to find the best suitable model for analysing students' classroom feedback. Many researchers have overlooked the preprocessing step in sentiment analysis. However this research has highlighted that the study

of preprocessing is an important step in any sentiment analysis study as it could increase or decrease the performance of the models. The study of different levels of preprocessing allows researchers in the sentiment analysis field to explore the best combination of preprocessing that is suitable for their application.

Another aspect we covered in this thesis is looking at less common features related to the domain and data source. While most of the researchers have explored the use of common features, our research highlighted that the use of additional features could increase the performance of the models noticeably. Therefore it is essential for researchers to explore other features in their studies to find their impact on their models.

In sentiment analysis, most of the studies used common machine learning techniques such as SVM, NB and ME. However our research has highlighted that using other classifiers could give a better performance. One example is the CNB classifier which has not been used much in the sentiment analysis research, however has performed extremely well with models with unbalanced classes such as models with the neutral class. Unbalanced classes is a general problem in sentiment analysis, therefore researchers could experiment with the CNB classifier to address this issue.

Our research has highlighted the importance of visualising the results of the analysis to make it useful to the users of the application. There is very little research about visualising sentiment analysis results and the focus on most of the studies has been towards the models themselves. Visualising the data in a format that is not understood by the user makes the analysis ineffective because the results are not usable. In our research, we identified that people of different genders could have different preferences for visualisation. This aspect deserves further investigation due to its potential effect on the usability of results.

Our research highlighted some aspects that we believe would be of interest to other researchers:

- In education, certain emotions are more relevant and frequent than others;
- Preprocessing has an important effect on classification performance; on our data we found that for polarity detection, higher preprocessing is useful, while for emotion detection, lower preprocessing is more appropriate;
- Visualisation preferences may vary by gender.

## 9.4 Future Work

Much research can be carried out in the future. In the polarity detection experiment, some of the preprocessing techniques increased the performance of models, while other decreased it.

Our research has shown the importance of investigating preprocessing further in sentiment analysis for other domains. In our study, a domain-based lexicon was used as a feature which increased the performance of the models. The lexicon's size was small as it was created from the collected students' feedback which was limited. Our findings can direct other researchers in other domains (e.g. tourism, products and movie) into using domain-based lexicons to increase the performance of their models. Muhammad et al. [116] research about Twitter sentiment analysis highlighted that using a domain-based lexicon in combination with a generic one can lead to a better model performance. The generic lexicon has the advantage of a wider term coverage, while the domain-based lexicon includes sentiment-bearing terms that may not be available in the generic one. Future work includes exploring the addition of a generic lexicon to the domain-based one. This could contribute to increasing the performance of the models.

In relation to the emotion detection, n-grams were the only common features that were explored. Future work includes exploring other common features such as POS-tagging which may increase the models' performance. In the other features experiments in both the polarity and emotion studies, the hashtag in combination with other features were found useful, however, its performance alone was not explored. It could be observed from the data, that students use hashtags to express sentiment. Therefore, hashtags will be explored further to find if alone they increase the performance of the models. Hashtags may be explored in other domains to investigate if they are used to express sentiment. Similar to the polarity, the addition of a generic lexicon in combination with the domain-based one may improve the performance of the models, therefore will be explored. Additionally, in this study the polarity and emotion labels were obtained from the students. For future work, an investigation of the relation between polarity and emotion can be done. This can be useful to acknowledge which learning emotions are associated with the different polarity classes.

In this study, less common features were explored. Some of these features led to a good performance with the models, for instance CNB performed well with datasets with unbalanced classes. This can be explored further in other domains to investigate whether these classifiers achieve a good performance.

In sarcasm detection, most of the previous studies were based on a large dataset, while we only used a small dataset due to the limited resources of data. It would be useful to apply this study to a larger dataset to explore the difference in performance. It would also be interesting to apply this study to other domains. In this study, human annotators were used to label the sarcastic sentences. For future work it would be useful to explore different methods to label them automatically. This can allow the implementation of the other features models to detect polarity and emotion.

As identified by this research, visualisation of the data is an important element in sentiment analysis. Most of the research in sentiment analysis has been related to the performance of the models and not on the visualisation of the models. In real-life applications, the visualisation of the data is needed therefore it should be investigated further. Our findings showed that gender had an effect on certain visualisations. Due to the small amount of participants, it is difficult to assess if this is a general effect, or just a characteristic of our sample. However, this could be an interesting area to investigate further. Additionally, exploring a larger dataset could be allow the investigation of demographic factors (e.g. age, country and department) further and their relation to the visualisation preferences.

As for the system evaluation, the sample of participants obtained was small, therefore, our findings were restricted. It would be beneficial to evaluate the system with a larger sample. Due to the low sample size, the demographic factors such as gender were not explored. It would be interesting to explore the impact demographic factors on the evaluation. Additionally, the system was only evaluated in the computing department, therefore, the results may be considered bias. Consequently it would be interesting to evaluate the system in other departments to explore the difference in evaluation.





# References

- [1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, pages 30–38, 2011.
- [2] Jean Aitchison. *Language change: progress or decay?* Cambridge University Press, 2001.
- [3] Nabeela Altrabsheh, Mihaela Cocea, and Sanaz Fallahkhair. Learning sentiment from students’ feedback for real-time interventions in classrooms. In *Adaptive and Intelligent Systems*, pages 40–49. Springer, 2014.
- [4] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Text, Speech and Dialogue*, pages 196–205. Springer, 2007.
- [5] Sarabjot S Anand and Alex G Büchner. *Decision support using data mining*. Financial Times Management, 1998.
- [6] Richard Anderson, Ruth Anderson, Beth Simon, Steven A. Wolfman, Tammy Van-DeGrift, and Ken Yasuhara. Experiences with a tablet pc based lecture presentation system in computer science courses. In *Proceedings of the 35th Special Interest Group on Computer Science Education Technical Symposium on Computer Science Education*, 4, pages 56–60. ACM, 2004.
- [7] Amanda L Andrei. Development and evaluation of tagalog linguistic inquiry and word count (liwc) dictionaries for negative and positive emotion. Technical report, The Mitre Corporation, 2014. URL [http://www.mitre.org/sites/default/files/publications/pr\\_14-3858-development-evaluation-of-tagalog-linguistic-inquiry.pdf](http://www.mitre.org/sites/default/files/publications/pr_14-3858-development-evaluation-of-tagalog-linguistic-inquiry.pdf).
- [8] Ivon Arroyo, David G. Cooper, Winslow Burleson, Beverly Park Woolf, Kasia Muldner, and Robert Christopherson. Emotion sensors go to school. In *Artificial Intelligence in Education*, volume 200 of *Frontiers in Artificial Intelligence and Applications*, pages 17–24. IOS Press, 2009.
- [9] Sitaram Asur, Bernardo Huberman, et al. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology*, volume 1, pages 492–499. IEEE, 2010.
- [10] Henning Bar, Erik Tews, and Guido Robling. Improving feedback and classroom interaction using mobile phones. *Proceedings of Mobile Learning*, pages 55–62, 2005.

- [11] Brijesh Kumar Baradwaj and Saurabh Pal. Mining educational data to analyze students' performance. *International Journal of Advanced Computer Science and Applications*, 2(6):63–69, 2011.
- [12] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *International Conference on Computational Linguistics*, volume 23 of 10, pages 36–44. Association for Computational Linguistics, 2010.
- [13] Scott Bateman, Carl Gutwin, and Miguel Nacenta. Seeing things in the clouds: the effect of visual features on tag cloud selections. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, HT '08, pages 193–202. ACM, 2008.
- [14] P Bhargavi and S Jyothi. Applying naive bayes data mining technique for classification of agricultural land soils. *International journal of computer science and network security*, 9(8):117–122, 2009.
- [15] Albert Bifet and Eibe Frank. Sentiment knowledge discovery in twitter streaming data. In *Discovery Science*, pages 1–15. Springer, 2010.
- [16] Haji Binali, Chen Wu, and Vidyasagar Potdar. Computational approaches for emotion detection in text. In *Digital Ecosystems and Technologies (DEST)*, pages 100–107. IEEE, 2010.
- [17] Zeynep Boynukalin. Emotion analysis of Turkish texts by using machine learning methods. Master's thesis, Middle East Technical University, 2012.
- [18] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [19] Gaya Buddhinath and Damien Derry. A simple enhancement to one rule classification. Technical report, University of Melbourne, Australia., Department of Computer Science & Software Engineering, 2006. URL <http://www.buddhinath.net/OtherLinks/Documents/Improved%20OneR%20Algorithm.pdf>.
- [20] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, pages 205–211. ACM, 2004.
- [21] E Holly Buttner. How do we “dis” students?: A model of (dis) respectful business instructor behavior. *Journal of Management Education*, 28(3):319–334, 2004.
- [22] Peter Cabena, Pablo Hadjinian, Rolf Stadler, Jaap Verhees, and Alessandro Zanasi. *Discovering data mining: from concept to implementation*. Prentice-Hall, Inc., 1998.
- [23] Jeff Cain, Esther P Black, and Jurgen Rohr. An audience response system strategy to improve student motivation, attention, and feedback. *American Journal of Pharmaceutical Education*, 73(2):1–7, 2009.
- [24] Toon Calders and Mykola Pechenizkiy. Introduction to the special section on educational data mining. *Association for Computing Machinery's Special Interest Group on Knowledge Discovery and Data Mining.*, 13(2):3–6, May 2012.

- [25] David Calleele, Eric Neufeld, and Kevin Schneider. Visualizing emotional requirements. In *Fourth International Workshop on Requirements Engineering Visualization*, pages 1–10. IEEE, 2009.
- [26] Soumaya Chaffar and Diana Inkpen. Using a heterogeneous dataset for emotion analysis in text. In *Advances in Artificial Intelligence*, pages 62–67. Springer, 2011.
- [27] Wilas Chamlertwat, Pattarasinee Bhattarakosol, Tippakorn Rungkasiri, and Choochart Haruechaiyasak. Discovering consumer insight from twitter via sentiment analysis. *Journal of Universal Computer Science*, 18(8):973–992, April 2012.
- [28] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear svm. *Journal of Machine Learning Research*, 11:1471–1490, Aug 2010.
- [29] Li Chen, Feng Wang, Luole Qi, and Fengfeng Liang. Experiment on sentiment embedded comparison interface. *Knowledge-Based Systems*, 64:44–58, 2014.
- [30] Yoonjung Choi and Janyce Wiebe. +/-effectwordnet: Sense-level lexicon acquisition for opinion inference. *Empirical Methods in Natural Language Processing*, pages 1181–1191, 2014.
- [31] Krzysztof J Cios, Anna Teresinska, Stefania Konieczna, Joanna Potocka, and Sunil Sharma. Diagnosing myocardial perfusion from pect bull’s-eye maps-a knowledge discovery approach. *IEEE Engineering in Medicine and Biology Magazine*, 19(4): 17–25, 2000.
- [32] W.B. Claster, M. Cooper, and P. Sallis. Thailand – tourism and conflict: Modeling sentiment from twitter tweets using naive bayes and unsupervised artificial neural nets. In *Second International Conference on Computational Intelligence, Modelling and Simulation*, pages 89–94, Sept 2010.
- [33] William S Cleveland and Robert McGill. Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833, 1985.
- [34] Monica F. Cox, Jeeyeon Hahn, Nathan McNeill, Osman Cekic, Jiabin Zhu, and Jeremi London. Enhancing the quality of engineering graduate teaching assistants through multidimensional feedback. *Advances in Engineering Education*, 2:1–20, 2011.
- [35] Marlena A Creusere. Theories of adults’ understanding and use of irony and sarcasm: Applications to and evidence from research with children. *Developmental Review*, 19 (2):213–262, 1999.
- [36] David Crystal. *English as a global language*. Cambridge University Press, 2012.
- [37] Hugh M Culbertson and Richard D Powers. A study of graph comprehension difficulties. *Educational Technology Research and Development*, 7(3):97–110, 1959.
- [38] S. Cummins, L. Burd, and A. Hatch. Using feedback tags and sentiment analysis to generate sharable learning resources investigating automated sentiment analysis of feedback tags in a programming course. *Advanced Learning Technologies (ICALT)*, 10:653–657, July 2010.

- [39] Bart Czernicki. Creating data visualizations for analysis. In *Silverlight 4 Business Intelligence Software*, pages 175–217. Springer, 2010.
- [40] Harry L Dangel and Charles Xiaoxue Wang. Student response systems in higher education: Moving beyond linear teaching and surface learning. *Journal of Educational Technology Development and Exchange*, 1(1):93–104, 2008.
- [41] Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *Convention Communication, Interaction and Social Intelligence*, volume 1, pages 53–59, 2008.
- [42] Dmitry Davidov, Oren Tsur, and Ari Rappoport. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, 10, pages 107–116. Association for Computational Linguistics, 2010.
- [43] Hugh C Davis, Kate Dickens, Manuel Leon Urrutia, Sánchez Vera, Maria del Mar, and Su White. Moocs for universities and learners an analysis of motivating factors, Apr 2014. URL <http://eprints.soton.ac.uk/363714/1/DavisEtAl2014MOOCsCSEDUFinal.pdf>.
- [44] R. De Groot. Data mining for tweet sentiment classification. Master’s thesis, Utrecht University, 2012.
- [45] Frank Dellaert, Thomas Polzin, and Alex Waibel. Recognizing emotion in speech. In *Fourth International Conference on Spoken Language*, volume 3, pages 1970–1973. IEEE, 1996.
- [46] Katherine J Denker. Student response systems and facilitating the large lecture basic communication course: Assessing engagement and learning. *Communication Teacher*, 27(1):50–69, 2013.
- [47] Sanjeev Dhawan, Kulvinder Singh, and Vandana Khanchi. A framework for polarity classification and emotion mining from text. *International Journal Of Engineering And Computer Science*, 3(8):7431–7436, 2004.
- [48] Amandeep Dhir, Khalid Buragga, and Abeer A Boreqqah. Tweeters on campus: Twitter a learning tool in classroom? *Journal of Universal Computer Science*, 19(5): 672–691, 2013.
- [49] Khaldoon Dhou. *Toward a better understanding of viewers’ perceptions of tag clouds: relative size judgment*. PhD thesis, UNC Charlotte, 2013.
- [50] Sidney D’mello and Art Graesser. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):23–39, 2012.
- [51] Sidney D’Mello, Tanner Jackson, Scotty Craig, et al. Autotutor detects and responds to learners affective and cognitive states. In *Workshop on Emotional and Cognitive Issues at the International Conference on Intelligent Tutoring Systems*, pages 31–43, 2008.

- [52] Linda S Elting, Charles G Martin, Scott B Cantor, and Edward B Rubenstein. Influence of data display formats on physician investigators' decisions to stop clinical trials: prospective trial with repeated measures. *British Medical Journal*, 318(7197):1527–1531, 1999.
- [53] Usama M Fayyad. Data mining and knowledge discovery: Making sense out of data. *IEEE Intelligent Systems*, 11(5):20–25, 1996.
- [54] Tian Feng, Liang Huijun, Li Longzhuang, and Zheng Qinghua. Sentiment classification in turn-level interactive Chinese texts of e-learning applications. In *IEEE 12th International Conference on Advanced Learning Technologies (ICALT)*, pages 480–484, July 2012.
- [55] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.
- [56] Michael Gamon. Sentiment classification on customer feedback data: Noisy data, large feature vectors, and the role of linguistic analysis. In *International Conference on Computational Linguistics*, volume 20, pages 841–847, 2004.
- [57] Michael Gamon, Anthony Aue, Simon Corston-Oliver, and Eric Ringger. Pulse: Mining customer opinions from free text. In *Proceedings of the Sixth International Conference on Advances in Intelligent Data Analysis*, volume 6 of 5, pages 121–132. Springer-Verlag, 2005.
- [58] Ramon Garrote and Tomas Pettersson. Lecturers' attitudes about the use of learning management systems in engineering education: A swedish case study. *Australasian Journal of Educational Technology*, 23(3):327–349, 2007.
- [59] Edward F Gehringer. Applications for supporting collaboration in the classroom. In *American Society for Engineering Education*. American Society for Engineering Education, 2012.
- [60] Nahum Gershon. Visualization of an imperfect world. *IEEE Computer Graphics and Applications*, 18(4):43–45, 1998.
- [61] Jan Geryk. Visual analytics by animations in higher education. In *European Conference on e-Learning*, pages 565–572. Academic Conferences International Limited, 2013.
- [62] Eric Gilbert and Karrie Karahalios. Widespread worry and the stock market. In *Association for the Advancement of Artificial Intelligence*, volume 4, pages 59–65, 2010.
- [63] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 2009. URL [http://s3.eddieoz.com/docs/sentiment\\_analysis/Twitter\\_Sentiment\\_Classification\\_using\\_Distant\\_Supervision.pdf](http://s3.eddieoz.com/docs/sentiment_analysis/Twitter_Sentiment_Classification_using_Distant_Supervision.pdf).
- [64] Alec Go, Lei Huang, and Richa Bhayani. Twitter sentiment analysis. CS224N Project Report, Stanford, 2009. URL <http://www-nlp.stanford.edu/courses/cs224n/2009/fp/3.pdf>.

- [65] B. Gokulakrishnan, P. Priyanthan, T. Ragavan, N. Prasath, and A. Perera. Opinion mining and sentiment analysis on a twitter data stream. *Advances in Information and communications technology for Emerging Regions*, 13:182–188, Dec 2012.
- [66] Yoav Goldberg and Michael Elhadad. SplitSVM: Fast, space-efficient, non-heuristic, polynomial kernel computation for NLP applications. *Annual Meeting of the Association of Computational Linguistics*, 26:237–240, 2008.
- [67] Roberto Gonzalez-Ibanez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*, pages 581–586. Association for Computational Linguistics, 2011.
- [68] Arthur Graesser, Patrick Chipman, Brandon King, Bethany McDaniel, and Sidney D’Mello. Emotions and learning with auto tutor. *Frontiers in artificial intelligence and application*, 158:117–139, 2007.
- [69] Michelle L Gregory, Nancy Chinchor, Paul Whitney, Richard Carter, Elizabeth Hetzler, and Alan Turner. User-directed sentiment analysis: Visualizing the affective content of documents. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 23–30. Association for Computational Linguistics, 2006.
- [70] Gabriela Grosseck and Carmen Holotescu. Can we use twitter for educational activities. In *The fourth international conference on eLearning and software for education*, volume 4, 2008. URL <https://adlunap.ro/else/papers/015.-697.1.Grosseck%20Gabriela-Can%20we%20use.pdf>.
- [71] Satarupa Guha, Aditya Joshi, Vasudeva Varma, and Hyderabad Gachibowli. Sentibase: Sentiment analysis in twitter on a budget. *Proceedings of the Ninth International Workshop on Semantic Evaluation* ., 9:590—594, 2015.
- [72] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4): 1334–1345, 2007.
- [73] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: an update. *ACM Association for Computing Machinery’s Special Interest Group on Knowledge Discovery and Data Mining*, 11(1):10–18, 2009.
- [74] Charles F Harrington, Scott A Gordon, and Timothy J Schibik. Course management system utilization and implications for practice: A national survey of department chairpersons. Technical Report 4, University of West Georgia Richards, 2004. URL <http://www.westga.edu/~distance/ojdla/winter74/harrington74.htm>.
- [75] Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007.
- [76] Shawndra Hill, Aman Nalavade, and Adrian Benton. Social tv: Real-time social media response to tv advertising. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, 12, pages 4–12. ACM, 2012.

- [77] Sarmin Hossain and Laurence Brooks. Fuzzy cognitive map modelling educational software adoption. *Computers & Education*, 51(4):1569–1588, 2008.
- [78] Diana Inkpen, Fazel Keshtkar, and Diman Ghazi. Analysis and generation of emotion in texts. *Knowledge engineering: principle and technique*, 2:3–14, 2009.
- [79] Siddharth Jain and Sushobhan Nayak. Sentiment analysis of movie reviews: A study of features and classifiers. CS221 Project Report, Stanford, 2013. URL [http://web.stanford.edu/~nayaks/reportsFolder/cs221\\_report.pdf](http://web.stanford.edu/~nayaks/reportsFolder/cs221_report.pdf).
- [80] Raquel Justo, Thomas Corcoran, Stephanie M. Lukin, Marilyn Walker, and M. Ines Torres. Extracting relevant knowledge for the detection of sarcasm and nastiness in the social web. *Knowledge-Based Systems*, 69:124–133, 2014.
- [81] Asha Gowda Karegowda, AS Manjunath, and MA Jayaram. Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, 2(2):271–277, 2010.
- [82] Walter Kasper and Mihaela Vela. Sentiment analysis for hotel reviews. In *Computational Linguistics-Applications Conference*, pages 45–52, 2011.
- [83] Zied Kechaou, M Ben Ammar, and Adel M Alimi. Improving e-learning with sentiment analysis of users’ opinions. *Global Engineering Education Conference*, 6: 1032–1038, 2011.
- [84] Fazel Keshtkar and Diana Inkpen. A corpus-based method for extracting paraphrases of emotion terms. In *Computational approaches to Analysis and Generation of emotion in Text*, pages 35–44. Association for Computational Linguistics, 2010.
- [85] Vinh Ngoc Khuc, Chaitanya Shivade, Rajiv Ramnath, and Jay Ramanathan. Towards building large-scale distributed systems for twitter sentiment analysis. In *Proceedings of the 27th ACM Symposium on Applied Computing, SAC '12*, pages 459–464. ACM, 2012.
- [86] Stephen Kinsella. Many to one: Using the mobile phone to interact with large classes. *British Journal of Educational Technology*, 40(5):956–958, 2009.
- [87] Claire Knight. Visualisation effectiveness. In *International Conference on Imaging Science, Systems, and Technology*, pages 249–254, 2001.
- [88] Kyle Koh, Bongshin Lee, Bohyoung Kim, and Jinwook Seo. Maniwordle: Providing flexible control over wordle. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1190–1197, 2010.
- [89] Barry Kort, Rob Reilly, and Rosalind W Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies*, pages 43–436. IEEE Computer Society, 2001.
- [90] Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.

- [91] Taku Kudo and Yuji Matsumoto. Chunking with support vector machines. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*, NAACL '01, pages 1–8. Association for Computational Linguistics, 2001.
- [92] Akshi Kumar and Teeja Mary Sebastian. Sentiment analysis on twitter. *IJCSI International Journal of Computer Science Issues*, 9(3):372–378, 2012.
- [93] Lukasz A Kurgan and Petr Musilek. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(01):1–24, 2006.
- [94] Chee Kian Leong, Yew Haur Lee, and Wai Keong Mak. Mining sentiments in sms texts for teaching evaluation. *Expert Systems with Applications*, 39(3):2584–2589, 2012.
- [95] Stephan Lewandowsky and Ian Spence. The perception of statistical graphs. *Sociological Methods & Research*, 18(2-3):200–242, 1989.
- [96] James R Lewis. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.
- [97] Yang Li, Bin-Xing Fang, You Chen, and Li Guo. A lightweight intrusion detection model based on feature selection and maximum entropy model. In *Communication Technology, 2006. ICCT'06. International Conference on*, pages 1–4. IEEE, 2006.
- [98] Shu-Sheng Liaw. Investigating students' perceived satisfaction, behavioral intention, and effectiveness of e-learning: A case study of the blackboard system. *Computers & Education*, 51(2):864–873, 2008.
- [99] CC Liebrecht, FA Kunneman, and APJ van den Bosch. The perfect solution for detecting sarcasm in tweets# not. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 29–37, 2013.
- [100] Diane J Litman and Kate Forbes-Riley. Predicting student emotions in computer-human tutoring dialogues. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 351–358. Association for Computational Linguistics, 2004.
- [101] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. Practical resources for assessing and reporting intercoder reliability in content analysis research projects, 2004. URL <http://astro.temple.edu/~lombard/reliability/>.
- [102] Bin Lu and Benjamin K Tsou. Combining a large sentiment lexicon and machine learning for subjectivity classification. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 6, pages 3311–3316. IEEE, 2010.
- [103] Arnold M Lund. Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6, 2001.



- [104] Teddy Mantoro, Media Ayu, Emir Habul, Ana U Khasanah, et al. Survnvote: A free web based audience response system to support interactivity in the classroom. In *Open Systems (ICOS), 2010 IEEE Conference on*, pages 34–39. IEEE, 2010.
- [105] Aaron Marcus. The ten commandments of color. *Computer Graphics Today*, 3(10): 7–14, 1986.
- [106] Adam Marcus, Michael S. Bernstein, Osama Badar, David R. Karger, Samuel Madden, and Robert C. Miller. Twitinfo: Aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, volume 11 of *CHI '11*, pages 227–236. ACM, 2011.
- [107] J.M. Martin, A. Ortigosa, and R.M. Carro. Sentbuk: Sentiment analysis for e-learning environments. In *International Symposium on Computers in Education (SIIE)*, volume 12, pages 212–217, Oct 2012.
- [108] Margie Martyn. Clickers in the classroom: An active learning approach. *Educause quarterly*, 30(2):71–74, 2007.
- [109] Kenji Mase. Recognition of facial expression from optical flow. *IEICE Transactions on Information and Systems*, 74(10):3474–3483, 1991.
- [110] Carmel McNaught and Paul Lam. Using wordle as a supplementary research tool. *The qualitative report*, 15(3):630–643, 2010.
- [111] Yelena Mejova. Sentiment analysis: An overview. comprehensive exam paper, 2009. URL <http://www.cs.uiowa.edu/~ymejova/publications/CompsYelenaMejova.pdf>.
- [112] Saif M Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.
- [113] Saif M Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. In *Second Joint Conference on Lexical and Computational Semantics (\*SEM)*, volume 2, pages 321–327, 2013.
- [114] Asmaa Mountassir, Houda Benbrahim, and Ilham Berrada. An empirical study to address the problem of unbalanced data sets in sentiment classification. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2012*, pages 3298–3303. IEEE, 2012.
- [115] K. Mouthami, K.N. Devi, and V.M. Bhaskaran. Sentiment analysis and classification based on textual reviews. In *Information Communication and Embedded Systems (ICICES)*, volume 1, pages 271–276, Feb 2013.
- [116] Aminu Muhammad, Nirmalie Wiratunga, Robert Lothian, and Richard Glassey. Domain-based lexicon enhancement for sentiment analysis. In *Proceedings of the British Computer Society's Specialist Group on Artificial Intelligence Workshop on Social Media Analysis*, volume 1110, pages 7–18, 2013.

- [117] Joseph M Mula and Marie Kavanagh. Click go the students, click-click-click: The efficacy of a student response system for engaging students to improve feedback and performance. *e-Journal of Business Education and Scholarship of Teaching*, 3(1): 1–17, 2009.
- [118] Myriam Munezero, Calkin Suero Montero, Maxim Mozgovoy, and Erkki Sutinen. Exploiting sentiment analysis to track emotions in students’ learning diaries. In *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*, pages 145–152. ACM, 2013.
- [119] Federico Neri, Carlo Aliprandi, Federico Capeci, Montserrat Cuadros, and Tomas By. Sentiment analysis on social media. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 919–926. IEEE Computer Society, 2012.
- [120] Kiyhoshi Nosu and Tomoya Kurokawa. A multi-modal emotion-diagnosis system to support e-learning. In *Innovative Computing, Information and Control*, volume 2, pages 274–278. IEEE, 2006.
- [121] Jeremy Novak and Michael Cowling. The implementation of social networking as a tool for improving student participation in the classroom. In *ISANA International Academy Association Conference Proceedings*, volume 22, pages 1–10, 2011.
- [122] Kerry O’Regan. Emotion and e-learning. *Journal of Asynchronous learning networks*, 7(3):78–92, 2003.
- [123] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541, 2014.
- [124] Sevgi Ozkan and Refika Koseler. Multi-dimensional students’ evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers & Education*, 53(4):1285–1296, 2009.
- [125] Neelamadhab Padhy, Dr Mishra, Rasmita Panigrahi, et al. The survey of data mining applications and feature scope. *International Journal of Computer Science, Engineering and Information Technology*, 2(3), 2012.
- [126] Alexander Pak and Patrick Paroubek. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, volume 5, pages 436–439, 2010.
- [127] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, volume 10, pages 1320–1326, 2010.
- [128] Georgios Paltoglou, Stéphane Gobron, Marcin Skowron, Mike Thelwall, and Daniel Thalmann. Sentiment analysis of informal textual communication in cyberspace. In *Engage 2010, Springer LNCS State-of-the-Art Survey*, pages 13–25, 2010.
- [129] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Annual Meeting on Association for Computational Linguistics*, 42:271–278, 2004.

- [130] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
- [131] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. *ACL-02 Conference on Empirical Methods in Natural Language Processing*, 10:79–86, 2002.
- [132] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.
- [133] John C. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods - Support Vector Learning*, pages 41–158. MIT Press, January 1998. URL <http://research.microsoft.com/apps/pubs/default.aspx?id=68391>.
- [134] J Poulis, C Massen, E Robens, and M Gilbert. Physics lecturing with audience paced feedback. *American Journal of Physics*, 66(5):439–441, 1998.
- [135] Ann Poulos and Mary Jane Mahony. Effectiveness of feedback: the students perspective. *Assessment & Evaluation in Higher Education*, 33(2):143–154, 2008.
- [136] Suhaas Prasad. Micro-blogging sentiment analysis using bayesian classification methods. CS224N Project Report, Stanford, 2010. URL <http://nlp.stanford.edu/courses/cs224n/2010/reports/suhaasp.pdf>.
- [137] Ashequl Qadir and Ellen Riloff. Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics*, pages 1203–1209, 2014.
- [138] Jason D Rennie, Lawrence Shih, Jaime Teevan, David R Karger, et al. Tackling the poor assumptions of naive bayes text classifiers. In *ICML*, volume 20, pages 616–623. Washington DC), 2003.
- [139] Daniela Retelny, Jeremy Birnholtz, and Jeffrey Hancock. Tweeting for class: using social media to enable student co-construction of lectures. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 203–206. ACM, 2012.
- [140] Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. Sarcasm as contrast between a positive sentiment and negative situation. In *EMNLP*, pages 704–714, 2013.
- [141] D Rogers, P Coughlan, and K Kearney. Playschool for postgrads: facilitating creativity and innovation in doctoral education using digital resources in, the digital learning revolution in ireland: Case studies in practice from ndlr. pages 165–179, 2013.
- [142] Cristóbal Romero, Sergio Gutiérrez, Manuel Freire, and Sebastián Ventura. Mining and visualizing visited trails in web-based educational systems. In *Educational Data Mining*, pages 182–186, 2008.

- [143] Shlomo Romi, Ramon Lewis, Joel Roache, and Philip Riley. The impact of teachers' aggressive management techniques on students' attitudes to schoolwork. *The Journal of Educational Research*, 104(4):231–240, 2011.
- [144] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Ninth International Conference on Language Resources and Evaluation*, page 810–817, 2014.
- [145] Matthias Schonlau and Ellen Peters. Graph comprehension: An experiment in displaying data as bar charts, pie charts and tables with and without the gratuitous 3rd dimension. *Social Science Research Network Working Paper Series*, pages 1–16, 2008.
- [146] Eusebio Scornavacca, Sid Huff, and Stephen Marshall. Mobile phones in the classroom: if you can't beat them, join them. *Communications of the ACM*, 52(4):142–146, 2009.
- [147] Priti Shah and Eric G Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, 3(3):560–578, 2011.
- [148] Priti Shah and James Hoeffner. Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69, 2002.
- [149] Nor Shahriza Abdul Karim, Siti Hawa Darus, and Ramlah Hussin. Mobile phone applications in academic library services: a students' feedback survey. *Campus-Wide Information Systems*, 23(1):35–51, 2006.
- [150] SG Shamay-Tsoory, Rachel Tomer, and Judith Aharon-Peretz. The neuroanatomical basis of understanding sarcasm and its relationship to social cognition. *Neuropsychology*, 19(3):288–300, 2005.
- [151] Tal Shany-Ur, Pardis Poorzand, Scott N. Grossman, Matthew E. Growdon, Jung Y. Jang, Robin S. Ketelle, Bruce L. Miller, and Katherine P. Rankin. Comprehension of insincere communication in neurodegenerative disease: Lies, sarcasm, and theory of mind. *Cortex*, 48(10):1329–1341, 2012.
- [152] C Shearer. The crisp-dm model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22, 2000.
- [153] Liping Shen, Minjuan Wang, and Ruimin Shen. Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Educational Technology & Society*, 12(2):176–189, 2009.
- [154] Wing So and Simon Wah. A study on the acceptance of mobile phones for teaching and learning with a group of pre-service teachers in hong kong. Technical Reports, 2008. URL <http://repository.ied.edu.hk/dspace/handle/2260.2/9992>.
- [155] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 151–161. Association for Computational Linguistics, 2011.

- [156] Dan Song, Hongfei Lin, and Zhihao Yang. Opinion mining in e-learning system. In *Network and Parallel Computing Workshops, 2007. NPC Workshops. IFIP International Conference on*, volume 6, pages 788–792, Sept 2007.
- [157] Jeffrey R Stowell and Jason M Nelson. Benefits of electronic audience response systems on student participation, learning, and emotion. *Teaching of psychology*, 34(4):253–258, 2007.
- [158] Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.
- [159] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2): 267–307, 2011.
- [160] Songbo Tan, Xueqi Cheng, Yuefen Wang, and Hongbo Xu. Adapting naive bayes to domain adaptation for sentiment analysis. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, pages 337–349, 2009.
- [161] Mohsen Tavakol and Reg Dennick. Making sense of cronbach’s alpha. *International journal of medical education*, 2:53–55, 2011.
- [162] Jaime Teevan, Daniel Liebling, Ann Paradiso, Carlos Garcia Jurado Suarez, Curtis von Veh, and Darren Gehring. Displaying mobile feedback during a presentation. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*, pages 379–382. ACM, 2012.
- [163] Zhi Teng, Fuji Ren, and Shingo Kuroiwa. Emotion recognition from text based on the rough set theory and the support vector machines. In *Natural Language Processing and Knowledge Engineering*, pages 36–41. IEEE, 2007.
- [164] Vasileios Terzis, Christos Moridis, and Anastasios Economides. Measuring instant emotions based on facial expressions during computer-based assessment. *Personal & Ubiquitous Computing*, 17(1):43–52, 2013.
- [165] Mike Thelwall. Heart and soul: Sentiment strength detection in the social web with sentistrength. *Proceedings of the CyberEmotions*, pages 1–14, 2013.
- [166] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, and Di Cai. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61:2544—2558, 2010.
- [167] Tun Thura Thet, Jin-Cheon Na, Christopher S.G. Khoo, and Subbaraj Shakthikumar. Sentiment analysis of movie reviews on discussion boards using a linguistic approach. In *Proceedings of the 1st International CIKM Workshop on Topic-sentiment Analysis for Mass Opinion, TSA ’09*, pages 81–84. ACM, 2009.

- [168] Feng Tian, Qinghua Zheng, Ruomeng Zhao, Tonghao Chen, and Xinyan Jia. Can e-learner's emotion be recognized from interactive Chinese texts? In *Computer Supported Cooperative Work in Design, 2009. CSCWD 2009. 13th International Conference on*, volume 13, pages 546–551, 2009.
- [169] Feng Tian, Qinghua Zheng, and Deli Zheng. Mining patterns of e-learner emotion communication in turn level of chinese interactive texts: Experiments and findings. In *Computer Supported Cooperative Work in Design (CSCWD)*, pages 664–670. IEEE, 2010.
- [170] Feng Tian, Pengda Gao, Longzhuang Li, Weizhan Zhang, Huijun Liang, Yanan Qian, and Ruomeng Zhao. Recognizing and regulating e-learners emotions based on interactive chinese texts in e-learning systems. *Knowledge-Based Systems*, 55: 148–164, 2014.
- [171] Sarah E Torok, Robert F McMorris, and Wen-Chi Lin. Is humor an appreciated teaching tool? perceptions of professors' teaching styles and use of humor. *College Teaching*, 52(1):14–20, 2004.
- [172] Tibor Toth. Graphing data for decision making. Technical Reports, 2006. URL <http://udspace.udel.edu/handle/19716/2666>.
- [173] Erik Tromp and Mykola Pechenizkiy. Rule-based emotion detection on social media: Putting tweets on plutchik's wheel. *Computing Research Repository*, 2014.
- [174] C. Troussas, M. Virvou, K. Junshean Espinosa, K. Llaguno, and J. Caro. Sentiment analysis of facebook statuses using naive bayes classifier for language learning. *Information, Intelligence, Systems and Applications*, 4:1–6, 2013.
- [175] Oren Tsur, Dmitry Davidov, and Ari Rappoport. Icwsm-a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, pages 162–169, 2010.
- [176] Hitoshi T Uchiyama, Daisuke N Saito, Hiroki C Tanabe, Tokiko Harada, Ayumi Seki, Kousaku Ohno, Tatsuya Koeda, and Norihiro Sadato. Distinction between the literal and intended meanings of sentences: a functional magnetic resonance imaging study of metaphor and sarcasm. *Cortex*, 48(5):563–583, 2012.
- [177] Mr Saif Vohra and Jay Teraiya. Applications and challenges for sentiment analysis: A survey. *International Journal of Engineering*, 2(2):1–5, 2013.
- [178] Daniel Voyer, Sophie-Hélène Thibodeau, and Breanna J Delong. Context, contrast, and tone of voice in auditory sarcasm perception. *Journal of psycholinguistic research*, pages 1–25, 2014.
- [179] Cheng Wang, Changqin Quan, and F. Ren. Maximum entropy based emotion classification of chinese blog sentences. In *Natural Language Processing and Knowledge Engineering (NLP-KE)*, pages 1–7, Aug 2010.

- [180] Wansen Wang and Jiabin Wu. Emotion recognition based on cso&svm in e-learning. In *Seventh International Conference on Natural Computation (ICNC)*, volume 7, pages 566–570, 2011.
- [181] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing twitter" big data" for automatic emotion identification. In *International Confernece on Social Computing (SocialCom) Privacy, Security, Risk and Trust (PASSAT)*, volume 4, pages 587–592. IEEE, 2012.
- [182] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68, 2010.
- [183] I. H. Witten, E. Frank, and A. Hall Mark. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.
- [184] Mohamed Yassine and Hazem Hajj. A framework for emotion mining from text in on-line social networks. In *Data Mining Workshops (ICDMW), 2010 IEEE International Conference on*, pages 1136–1142. IEEE, 2010.
- [185] Kuat Yessenov and Saša Misailovic. Sentiment analysis of movie review comments. *Methodology*, pages 1–17, 2009.
- [186] Yasuhisa Yoshida, Tsutomu Hirao, Tomoharu Iwata, Masaaki Nagata, and Yuji Matsumoto. Transfer learning for multiple-domain sentiment analysis-identifying domain dependent/independent word polarity. In *AAAI Conference on Artificial Intelligence*, volume 25, pages 1286–1291, 2011.
- [187] Zheng Yuan and Matthew Purver. Predicting emotion labels for chinese microblog texts. In *SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data*, volume 40, 2012.
- [188] Jeff Zacks and Barbara Tversky. Bars and lines: A study of graphic communication. *Memory and Cognition*, 27:1073–1079, 1999.
- [189] Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. Technical Reports, June 2011. URL <http://www.hpl.hp.com/techreports/2011/HPL-2011-89.html>.
- [190] Siqi Zhao, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. Analyzing twitter for social tv: Sentiment extraction for sports. In *in Proceedings of the 2nd International Workshop on Future of Television*, volume 2, pages 11–18, 2011.







# Appendix A

## Ethical Approval



### Certificate of Fast Track Ethics Review

Project Title:	Sentiment analysis for students classroom feedback
Student Number:	618634
Application Date:	18/09/2014 15:06:37

You must download your referral certificate, print a copy and keep it as a record of this review.

You should submit your certificate to your FEC representative for further review.

The FEC representative for the School of Computing is Carl Adams

It is your responsibility to follow the University Code of Practice on Ethical Standards and any Department/School or professional guidelines in the conduct of your study including relevant guidelines regarding health and safety of researchers including the following:

- University Policy
- Safety on Geological Fieldwork

It is also your responsibility to follow University guidance on Data Protection Policy:

- General guidance for all data protection issues
- University Data Protection Policy

ProjectTitle:  
Sentiment analysis for students classroom feedback

SchoolOrDepartment:  
SOC

PrimaryRole:  
PostgraduateStudent

SupervisorName:  
Mihaela Cocea

HumanParticipants:  
Yes

HumanParticipantsWarning  
ParticipantInformationSheets:  
The participants are informed that the feedback is optional .The participants are informed the reasons of the data collection and the study.

ParticipantConfidentiality:  
The participants names are disposed of if given,however, they are asked to not give their names. The data is kept only in the access of the researcher (Me). The data is not distributed to anyone or published.

InvolvesNHSPatientsOrStaff:

No

NoConsentOrDeception:

No

InvolvesUninformedOrDependents:

No

DrugsPlacebosOrOtherSubstances:

No

BloodOrTissueSamples:

No

PainOrMildDiscomfort:

No

PsychologicalStressOrAnxiety:

No

ProlongedOrRepetitiveTesting:

No

FinancialInducements:

No

PhysicalEcologicalDamage:

No

HistoricalOrCulturalDamage:

No

HarmToAnimal:

No

HarmfulToThirdParties:

No

### Supervisor Review

Supervisor signature:

Date:

### Review by FEC Representative

Name of representative:

Comments:

Representative signature:

Date:





## Appendix B

# Feature Usefulness Experiment

With  
Neutral  
Preprocessing P1

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
UNI+2	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
	Negative-lexicon	Negative-lexicon	Embarrassed	Happy	Engagement
	Excitement	Happy	Excitement	Negative-lexicon	Enthusiasm
UNI+7	Neutral-lexicon	Neutral-lexicon	Engagement	Frustration	Bored
	Happy	Excitement	Positive-lexicon	Sarcasm	# of punctuation
	# of hashtags	# of hashtags	Negative-lexicon	Positive-lexicon	Negative-lexicon
	Frustration	Frustration	Frustration	Engagement	Appreciation
	Sarcasm	Sarcasm	Neutral-lexicon	Neutral-lexicon	Satisfaction
UNI+15	Engagement	Engagement	Confusion	Enthusiasm	Beginning-lecture
	Enthusiasm	Embarrassed	Sarcasm	# of hashtags	Sarcasm
	Embarrassed	Enthusiasm	Enthusiasm	Confusion	Anxiety
	Confusion	Confusion	# of hashtags	Anxiety	Awkward
	Anxiety	Anxiety	Anxiety	Embarrassed	Dissapointed
	Shame	Motivated	Awkward	# of punctuation	Confusion
	Awkward	Proud	Appreciation	Appreciation	Embarrassed
	Dissapointed	Satisfaction	Dissapointed	Relief	Middle-lecture
	Motivated	Relief	# of punctuation	Proud	End-lecture
	Relief	Shame	Satisfaction	Shame	Uninterested
	Proud	Appreciation	Uninterested	Satisfaction	# of emoticons
	Satisfaction	Dissapointed	Shame	Motivated	Shame
	Appreciation	Awkward	Relief	Awkward	Frustration
	# of question and exclamation marks	Uninterested	# of emoticons	Dissapointed	Motivated
	# of emoticons	Middle-lecture	Middle-lecture	Beginning-lecture	Proud
	Middle-lecture	End-lecture	End-lecture	# of emoticons	Relief

UNI+32	End-lecture	# of emoticons	# of question and exclamation marks	Uninterested	Neutral-lexicon
	# of punctuation	# of question and exclamation marks	Motivated	# of question and exclamation marks	# of question and exclamation marks
	Uninterested	# of punctuation	Proud	End-lecture	Positive-lexicon
	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	# of hashtags

### Preprocessing P2

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
UNI+2	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
UNI+7	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
	Negative-lexicon	Negative-lexicon	Embarrassed	Happy	Engagement
	Neutral-lexicon	Neutral-lexicon	Excitement	Positive-lexicon	Enthusiasm
	Excitement	Happy	Engagement	Negative-lexicon	Negative-lexicon
	Happy	Excitement	Positive-lexicon	Frustration	Bored
UNI+15	# of hashtags	# of hashtags	Frustration	Neutral-lexicon	Positive-lexicon
	Frustration	Frustration	Negative-lexicon	Sarcasm	# of punctuation
	Sarcasm	Sarcasm	Neutral-lexicon	Engagement	Appreciation
	Engagement	Engagement	Confusion	Enthusiasm	Satisfaction
	Enthusiasm	Embarrassed	Sarcasm	# of hashtags	Beginning-lecture
	Embarrassed	Enthusiasm	Enthusiasm	Confusion	Sarcasm
	Confusion	Confusion	# of hashtags	Anxiety	Proud
	Anxiety	Anxiety	Anxiety	Embarrassed	Relief
	Appreciation	Appreciation	Appreciation	# of punctuation	Motivated
	Awkward	Awkward	Awkward	Appreciation	Embarrassed
	Dissapointed	Dissapointed	Dissapointed	Relief	Frustration
	Motivated	Motivated	Motivated	Proud	End-lecture
	Proud	Proud	Proud	Shame	Middle-lecture
	# of emoticons	# of emoticons	# of emoticons	Satisfaction	# of emoticons

	Middle-lecture	Middle-lecture	Middle-lecture	Motivated	Shame
	End-lecture	End-lecture	End-lecture	Awkward	Uninterested
	# of question and exclamation marks	# of question and exclamation marks	# of question and exclamation marks	Dissapointed	Anxiety
	# of punctuation	# of punctuation	# of punctuation	Beginning-lecture	Dissapointed
	Uninterested	Uninterested	Uninterested	# of emoticons	Confusion
	Relief	Relief	Relief	Uninterested	Awkward
	Satisfaction	Satisfaction	Satisfaction	# of question and exclamation marks	Neutral-lexicon
	Shame	Shame	Shame	End-lecture	# of question and exclamation marks
<b>UNI+32</b>	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	# of hashtags

### Preprocessing P3

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
<b>UNI+2</b>	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
	Negative-lexicon	Negative-lexicon	Embarrassed	Positive-lexicon	Engagement
	Excitement	Happy	Excitement	Happy	Enthusiasm
	Neutral-lexicon	Neutral-lexicon	Engagement	Neutral-lexicon	Bored
<b>UNI+7</b>	Happy	Excitement	Positive-lexicon	Frustration	Neutral-lexicon
	# of hashtags	# of hashtags	Frustration	Sarcasm	Positive-lexicon
	Frustration	Frustration	Neutral-lexicon	Negative-lexicon	# of punctuation
	Sarcasm	Sarcasm	Negative-lexicon	Engagement	Appreciation
	Engagement	Engagement	Confusion	Enthusiasm	Negative-lexicon
	Enthusiasm	Embarrassed	Sarcasm	# of hashtags	Satisfaction



UNI+15	Embarrassed	Enthusiasm	Enthusiasm	Confusion	Beginning-lecture
	Confusion	Confusion	# of hashtags	Anxiety	Sarcasm
	Anxiety	Anxiety	Anxiety	Embarrassed	Confusion
	Dissapointed	Dissapointed	Dissapointed	# of punctuation	Embarrassed
	Motivated	Motivated	Motivated	Appreciation	Dissapointed
	Awkward	Awkward	Awkward	Relief	Awkward
	Appreciation	Appreciation	Appreciation	Proud	Anxiety
	Proud	Proud	Proud	Shame	Frustration
	Relief	Relief	Relief	Satisfaction	Motivated
	Satisfaction	Satisfaction	Satisfaction	Motivated	Proud
	# of emoticons	# of emoticons	# of emoticons	Awkward	Relief
	End-lecture	End-lecture	End-lecture	Dissapointed	End-lecture
	Middle-lecture	Middle-lecture	Middle-lecture	Beginning-lecture	Middle-lecture
	# of question and exclamation marks	# of question and exclamation marks	# of question and exclamation marks	# of emoticons	# of emoticons
	Shame	Shame	Shame	Uninterested	Uninterested

UNI+32	# of punctuation	# of punctuation	# of punctuation	# of question and exclamation marks	Shame
	Uninterested	Uninterested	Uninterested	End-lecture	# of question and exclamation marks
	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	# of hashtags

### Without

#### Preprocessing P1

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
UNI+2	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
	Negative-lexicon	Negative-lexicon	Embarrassed	Negative-lexicon	Engagement
	Neutral-lexicon	Neutral-lexicon	Excitement	Happy	Enthusiasm

UNI+7	Excitement	Excitement	Engagement	Frustration	Negative-lexicon
	Happy	Happy	Positive-lexicon	Positive-lexicon	Bored
	Frustration	Frustration	Negative-lexicon	Sarcasm	Appreciation
	# of hashtags	Sarcasm	Frustration	Engagement	Satisfaction
	Sarcasm	# of hashtags	Neutral-lexicon	Neutral-lexicon	Uninterested
	Engagement	Engagement	Confusion	# of hashtags	# of punctuation
	Enthusiasm	Embarrassed	Sarcasm	Enthusiasm	Beginning-lecture
UNI+15	Confusion	Enthusiasm	Anxiety	Confusion	# of emoticons
	Anxiety	Confusion	Enthusiasm	Anxiety	Middle-lecture
	Embarrassed	Anxiety	# of hashtags	Embarrassed	End-lecture
	Appreciation	Appreciation	Appreciation	# of punctuation	Motivated
	Awkward	Awkward	Awkward	Appreciation	Shame
	Motivated	Motivated	Motivated	Proud	Anxiety
	Dissapointed	Dissapointed	Dissapointed	Relief	Proud
	Proud	Proud	Proud	Satisfaction	Dissapointed
	Middle-lecture	Middle-lecture	Middle-lecture	Dissapointed	Confusion
	End-lecture	End-lecture	End-lecture	Motivated	Awkward
UNI+32	# of question and exclamation marks	# of question and exclamation marks	# of question and exclamation marks	Awkward	Relief
	# of emoticons	# of emoticons	# of emoticons	Shame	Embarrassed
	# of punctuation	# of punctuation	# of punctuation	Beginning-lecture	Frustration
	Relief	Relief	Relief	Uninterested	Sarcasm
	Satisfaction	Satisfaction	Satisfaction	# of question and exclamation marks	# of question and exclamation marks
	Shame	Shame	Shame	End-lecture	# of hashtags
	Uninterested	Uninterested	Uninterested	# of emoticons	Neutral-lexicon
UNI+32	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	Positive-lexicon

Preprocessing P2

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
UNI+2	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
UNI+7	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
	Neutral-lexicon	Neutral-lexicon	Embarrassed	Positive-lexicon	Engagement
	Negative-lexicon	Negative-lexicon	Excitement	Happy	Positive-lexicon
	Excitement	Excitement	Engagement	Negative-lexicon	Enthusiasm
	Happy	Happy	Positive-lexicon	Frustration	Negative-lexicon
	Frustration	Frustration	Frustration	Neutral-lexicon	Bored
	# of hashtags	Sarcasm	Neutral-lexicon	Sarcasm	Appreciation
UNI+15	Sarcasm	# of hashtags	Negative-lexicon	Engagement	Satisfaction
	Engagement	Engagement	Confusion	# of hashtags	Awkward
	Enthusiasm	Embarrassed	Sarcasm	Enthusiasm	Confusion
	Confusion	Enthusiasm	Anxiety	Confusion	Dissapointed
	Anxiety	Confusion	Enthusiasm	Anxiety	Embarrassed
	Embarrassed	Anxiety	# of hashtags	Embarrassed	Frustration
	Proud	Awkward	Proud	# of punctuation	Anxiety
	Motivated	Motivated	Motivated	Appreciation	Motivated
	Appreciation	Dissapointed	Relief	Proud	Proud
	Dissapointed	Appreciation	Dissapointed	Relief	# of emoticons
	Awkward	Proud	Satisfaction	Satisfaction	Middle-lecture
	Relief	# of emoticons	Shame	Dissapointed	End-lecture
	# of punctuation	# of punctuation	Appreciation	Motivated	Relief
	# of emoticons	Middle-lecture	Awkward	Awkward	Shame
	Middle-lecture	End-lecture	Uninterested	Shame	# of punctuation
	# of question and exclamation marks	# of question and exclamation marks	# of emoticons	Beginning-lecture	Uninterested
	End-lecture	Satisfaction	Middle-lecture	Uninterested	Sarcasm
	Satisfaction	Relief	End-lecture	# of question and exclamation marks	Beginning-lecture

UNI+32	Uninterested	Uninterested	# of question and exclamation marks	End-lecture	# of question and exclamation marks
	Shame	Shame	# of punctuation	# of emoticons	# of hashtags
	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	Neutral-lexicon

## Preprocessing P3

	ChiSquare	InfoGain	GainRatio	Correlation	OneR
UNI+2	Bored	Bored	Happy	Bored	Amused
	Amused	Amused	Bored	Amused	Excitement
	Positive-lexicon	Positive-lexicon	Amused	Excitement	Happy
UNI+7	Negative-lexicon	Negative-lexicon	Embarrassed	Positive-lexicon	Negative-lexicon
	Neutral-lexicon	Neutral-lexicon	Excitement	Happy	Engagement
	Excitement	Excitement	Engagement	Frustration	Enthusiasm
	Happy	Happy	Positive-lexicon	Neutral-lexicon	Bored
	Frustration	Frustration	Frustration	Negative-lexicon	Positive-lexicon
	# of hashtags	Sarcasm	Neutral-lexicon	Sarcasm	Neutral-lexicon
	Sarcasm	# of hashtags	Confusion	Engagement	Appreciation
	Engagement	Engagement	Negative-lexicon	# of hashtags	Satisfaction
	Enthusiasm	Embarrassed	Sarcasm	Enthusiasm	Anxiety
	Confusion	Enthusiasm	Anxiety	Confusion	Embarrassed
UNI+15	Anxiety	Confusion	Enthusiasm	Anxiety	Frustration
	Embarrassed	Anxiety	# of hashtags	Embarrassed	Dissapointed
	# of punctuation	# of punctuation	# of punctuation	# of punctuation	Confusion
	Uninterested	Uninterested	Uninterested	Appreciation	Awkward
	Shame	Shame	Relief	Proud	Motivated
	Relief	Relief	Satisfaction	Relief	Proud
	Proud	Proud	Shame	Satisfaction	Relief
	Satisfaction	Satisfaction	# of question and exclamation marks	Dissapointed	End-lecture

	# of question and exclamation marks	# of question and exclamation marks	Middle-lecture	Motivated	# of emoticons
	Middle-lecture	Middle-lecture	# of emoticons	Awkward	Middle-lecture
	# of emoticons	# of emoticons	End-lecture	Shame	# of punctuation
	End-lecture	End-lecture	Appreciation	Beginning-lecture	Shame
	Motivated	Motivated	Proud	Uninterested	Uninterested
	Dissapointed	Dissapointed	Motivated	# of question and exclamation marks	Sarcasm
	Appreciation	Appreciation	Dissapointed	End-lecture	Beginning-lecture
	Awkward	Awkward	Awkward	# of emoticons	# of question and exclamation marks
UNI+32	Beginning-lecture	Beginning-lecture	Beginning-lecture	Middle-lecture	# of hashtags





## Appendix C

# Emotion Feature Usefulness Experiment

### Emotions

ChiSquare InfoGain GainRatio CorrelationOneR

Polarity=pc Polarity=pc Polarity=pc Polarity=pc Polarity=positive

Polarity=ne Polarity=ne Polarity=ne Polarity=ne Polarity=negative

NoofHas NoofHas NoofHas NoofHas student\_sarcasm

NoofPun NoofPun NoofPun student\_sa NoofQEmarks

Polarity=ne Polarity=ne Polarity=ne NoofQEma NoofPun

NoofEmoti NoofEmoti NoofEmoti Polarity=ne NoofEmoti

NoofQEma NoofQEma NoofQEma Time=M Time=M

Time=B Time=B Time=B Time=B Polarity=neutral

student\_sa student\_sa student\_sa Time=E Time=B

Time=E Time=E Time=E NoofPun Time=E

Time=M Time=M Time=M NoofEmoti NoofHas

### 2 Emotions

ChiSquare InfoGain GainRatio CorrelationOneR

Polarity=pc Polarity=pc Polarity=pc Polarity=pc Time=B

Polarity=ne Polarity=ne Polarity=ne Polarity=ne NoofEmoti

NoofHas NoofHas student\_sa NoofHas NoofQEmarks

student\_sa student\_sa NoofHas student\_sa Time=E

NoofPun NoofPun NoofPun NoofQEma Time=M

Polarity=ne Polarity=ne Polarity=ne Polarity=ne Polarity=neutral

NoofQEma NoofQEma NoofQEma Time=B Polarity=negative

NoofEmoti NoofEmoti NoofEmoti Time=M NoofPun

Time=B Time=B Time=B NoofPun student\_sarcasm

Time=M Time=M Time=M NoofEmoti Polarity=positive

Time=E Time=E Time=E Time=E NoofHas

### Amused

ChiSquare InfoGain GainRatio CorrelationOneR

Polarity=pc Polarity=pc Polarity=pc Polarity=pc student\_sarcasm

Polarity=ne Polarity=ne Polarity=ne Polarity=ne Polarity=neutral

Polarity=ne Polarity=ne Polarity=ne NoofHas Polarity=negative

NoofHas NoofHas NoofHas Polarity=ne Time=E

NoofPun NoofPun NoofPun NoofPun NoofPun

student\_sa student\_sa student\_sa NoofQEma NoofQEmarks

NoofQEma NoofQEma NoofQEma student\_sa NoofEmoti

NoofEmoti NoofEmoti NoofEmoti Time=M Time=B

Time=B Time=B Time=B NoofEmoti Time=M

Time=M Time=M Time=M Time=E Polarity=positive

Time=E Time=E Time=E Time=B NoofHas

### Bored

ChiSquare InfoGain GainRatio CorrelationOneR

Polarity=pc Polarity=pc Polarity=pc Polarity=pc student\_sarcasm

Polarity=ne Polarity=ne Polarity=ne Polarity=ne Polarity=neutral

NoofHas student\_sa student\_sa student\_sa Polarity=negative

student\_sa NoofHas NoofHas NoofHas Time=E



NoofPun NoofPun NoofPun Time=B NoofPun  
 Polarity=ne Polarity=ne Polarity=ne Time=M NoofHas  
 NoofQEmas NoofQEmas NoofQEmas NoofQEmas NoofQEmas  
 NoofEmo NoofEmo NoofEmo Time=E Time=B  
 Time=B Time=B Time=B Polarity=ne Time=M  
 Time=M Time=M Time=M NoofPun Polarity=positive  
 Time=E Time=E Time=E NoofEmo NoofEmo

Excitement

ChiSquare InfoGain GainRatio CorrelationOneR  
 Polarity=pc Polarity=pc Polarity=pc Polarity=pc student\_sarcasm  
 Polarity=ne Polarity=ne Polarity=ne Polarity=ne Time=E  
 NoofHas NoofHas NoofHas NoofHas Polarity=neutral  
 Time=E Time=E Time=M Time=E Polarity=negative  
 student\_sa student\_sa Time=B Polarity=ne Time=M  
 Polarity=ne Polarity=ne Time=E Time=M NoofPun  
 Time=M Time=M Polarity=ne student\_sa NoofHas  
 Time=B Time=B NoofQEmas NoofPun NoofQEmas  
 NoofEmoti NoofEmoti NoofEmoti NoofEmoti NoofEmoti  
 NoofPun NoofPun NoofPun NoofQEmas Time=B  
 NoofQEmas NoofQEmas student\_sa Time=B Polarity=positive